Kongeriget Danmark

Patent application No.: PA 2002 01685

Date of filing: 01 November 2002

Applicant:
(Name and address)
Torben F. Ørntoft
Klinisk Biokemisk Afdeling
Skejby Sygehus
DK-8200 Aarhus N
Denmark

Title: Identifying distinct classes of bladder cancer.

IPC: -

This is to certify that the attached documents are exact copies of the above mentioned patent application as originally filed.

Patent- og Varemærkestyrelsen
Økonomi- og Erhvervsministeriet

20 November 2003

John Nielsen

Abstract

Bladder cancer is a common malignant disease characterised by frequent recurrences[1,2]. Important factors determining the disease course of the individual patient are the stage of disease at diagnosis and the presence of surrounding carcinoma *in situ*[3]. Despite significant efforts, no accepted immunohistological or molecular markers define clinically relevant subsets of bladder cancer. Here we report the identification of clinically relevant subclasses of bladder carcinoma using expression microarray analysis of 40 well-characterised bladder tumours. Hierarchical cluster analysis identified the three major stages (Ta, T1 and T2-4) and the Ta tumours were furthermore separated into well defined subgroups. We built a 32 gene molecular classifier using a cross validation approach, which classified benign and muscle invasive tumours with close correlation to pathological staging. The classifier provided new predictive information on disease progression in Ta tumours (*P*<0.005). Other classifiers contained up to 320 genes and had similar good performance. To delineate non-recurring Ta tumours from frequently recurring Ta tumours we analysed expression patterns in 31 tumours by applying a supervised learning classification methodology, which classified 75% of the samples correctly (*P*<0.006). Furthermore, gene expression profiles characterising each stage and subtype demonstrated their biological properties and form new potential targets for therapy.

Introduction

Bladder cancer in the form of transitional cell carcinomas is a common malignant disease characterized by frequent recurrences. An important factor determining the disease course of the patient is the stage of disease at diagnosis. Patients presenting with relatively harmless stage Ta superficial papillomas will have recurrences in 50% of cases but less than 10% will later on develop an invasive tumor. On the other hand the tumors that show a superficial invasion into submucosa, stage T1, have a recurrence rate of 70% and 30% of those patients will later develop a muscle invasive tumor. Finally, about 25% of patients present with an invasive stage T2 –4 tumor at diagnosis [1]. Another epithelial abnormality influencing the disease course is the possible presence of dysplasia or carcinoma in situ in the mucosa surrounding the tumor. Patient having such field disease have much more frequent recurrences and a relatively poor prognosis, as 37% die within 10 years [2].

DNA fingerprinting as well as comparative genomic hybridization (CGH) have demonstrated that metachronous bladder tumors are of the same clonal origin [3,4].However, it is still not understood how a stage T1 tumor in the left side of the bladder mucosa can share clonal origin with a stage T2 tumor occurring in the right side after a purported tumor free interval of more than one year. Theories on implantation or seeding of tumor cells exist but have never been proved[4]. CGH technology has also shown that the superficial stage Ta and mucosa invasive T1 tumors, although they microscopically may look similar, have a quite different chromosomal integrity [5]. Stage T1 tumors show many more losses and gains of chromosomal materials than do stage Ta tumors. This has led to the suggestion that stage Ta and stage T1 tumors represent clinically different diseases[6].

Recent advances in microarray technology have made it possible to characterize cancers based on the expression of thousands of genes. Parallel gene expression monitoring is a powerful tool for the analysis of the relation between tumors, for discovering new tumor subgroups (class discovery), for assigning tumors to pre-defined classes (class prediction), and for identifying co-regulated or tumor stage specific genes[7-11]. In a recent study of bladder cancer, we demonstrated functional groups of genes whose co-regulation formed the basis for separating bladder tumors into superficial and muscle invasive tumors [12].

Here we used microarrays with approximate 5000 full-length genes to analyze gene expression in 40 bladder tumors selected from a very large clinical specimen bank holding more than 35.000 samples from bladder cancer patients, prospectively followed for up to six years. The selection was based on the disease course, stage, grade, concomitant carcinoma in situ, and recurrence frequency, in such a way that the selected tumors represent the spectrum from harmless stage Ta grade 2 superficial papillomas to muscle invasive stage T2 grade 4 tumors. Our data demonstrate a distinctly different gene expression in Ta tumors that separate these into three groups, relatively harmless Ta grade 2 tumors, frequently recurring stage Ta grade 3 tumors, and stage Ta grade 3 tumors with surrounding carcinoma in situ that cluster together with the invasive tumors. The arrays identified even minor histological alterations as the presence of areas of squamous metaplasia in invasive tumors, or the presence of carcinoma in situ.. Co-regulated groups of genes, such as genes related to proliferation, immune response and transcription, being up- or down regulated at certain stages and grades, describe the cell biological events that characterize each of the clinically well-known bladder tumor stages. Finally, from a set of 30 to 320 classifying genes we classified the tumor samples with close correlation to the pathological staging, plus obtained additional information on progression of disease and recurrence of tumors, as well as presence of cacinoma in situ.

Results

From our bladder cancer specimen bank we selected tumors of different histological stages and grades from six groups of patients (Table 1): (a) 5 patients with pT_a grade II tumors (no recurrence); (b) 5 patients with pT_a grade III tumors (no prior pT_1 tumor or CIS); (c) 5 patients with pT_a grade III tumors (CIS but no prior pT_1 tumor); (d) 4 patients with pT_a grade III tumors (a prior pT_1 tumor and CIS); (e) 11 patients with pT_1 grade III tumors (no prior pT_{2+} tumor); and (f) 10 patients with primary invasive pT_{2+} grade III/IV tumors. See Supplementary Information; Table 1 for complete disease course. In total 40 preparations of RNA from tumor and 4 from normal urothelial tissue were labeled and hybridized to Affymetrix oligonucleotide microarrays with approximately 5000 full-length genes. Scanning identified the expression level of the genes utilizing antibody amplification of weakly expressed genes. Genes that did not vary throughout the data-set as e.g.

housekeeping genes were eliminated, and only the 1767 genes (26 %) that showed an expression level change in tumor tissue compared to normal urothelium were subjected to cluster analysis.

Sample clustering

A two-way hierarchical clustering of the tumor samples based on the 1767 gene-set remarkably separated all 40 tumors according to stages and grades with only few exceptions (Fig.1a). Two main branches holding the superficial pTa tumors and the invasive pT1 and pT2+ tumors, respectively, were identified. In the superficial branch two sub-clusters of tumors could be identified, one holding 8 tumors that had frequent recurrences and one holding 3 out of the five pTa grade 2 tumors with no recurrence. In the invasive branch it was remarkable to find four pTa grade 3 tumors clustering tightly with the muscle invasive T2 tumors. These pTa tumors showed concomitant carcinoma in situ in the surrounding mucosa. This indicates that this sub-fraction of pTa tumors have some of the more aggressive features found in muscle invasive tumors. The pT1 cluster could be separated into three sub-clusters one holding four tumors including a pTa tumor, of whom 2 had CIS, and two others with no clear clinical difference. The one stage pT1 grade 3 tumor that clustered with the stage pT2+ muscle invasive tumors was the only T1 tumor that showed a solid growth pattern, the other were papillomas. Nine out of ten pT2+ tumors were found in one single cluster. As another technique to demonstrate the remarkable separation of the tumors we used multidimensional scaling analysis (Fig. 1c).

In an attempt to reduce the number of genes needed for class prediction we identified those genes that were scored by the Cancer Genome Anatomy Project as belonging to cancer-related groups such as tumor suppressors, oncogenes, genes involved in DNA-damage, angiogenesis, apoptosis, cell cycle, cell behavior, cell signaling, development, gene regulation, and transcription. These genes were then isolated from the initial 1767 gene-set and those 88, which showed largest variation (SD of the gene vector >=4), were used for hierarchical clustering of the tumor samples. This gene-set of only 88 genes was able to identify the clinically relevant groups almost as exact as the 1767 gene-set (Fig.1b). This finding emphasizes that the tumor clustering is not simply reflecting larger amounts of stromal components in the invasive tumor biopsies. The frequently recurring Ta grade 3 tumors clustered two by two in four separate clusters. The four pTa grade 3 tumors surrounded by CIS were still located inside the invasive branch. One Ta tumor (1166-1) that clustered as a T1 tumor using 1767 genes repeated this position with the small 88 gene-set. It cannot be ruled out that this tumor in reality is a T1 grade 3 tumor.

Gene clustering

Hierarchical cluster analysis of the 1767 genes revealed several characteristic profiles in which there was a distinct difference between the tumor groups (Fig. 1 d, black lines identifying clusters A to J).

Cluster A contains genes that show low expression in normal urothelium and stage T1 tumors, a medium level in stage T2 and a very high level in all the Ta grade 3 tumors (Fig.2a). This cluster contains 8 transcription factors as well as other nuclear genes related to transcriptional activity (See Supplementary Information; Figure 1 & 2 for enlarged views of cluster A-J). The high transcriptional activity may be related to both a high metabolic activity as well as an increased cell proliferation. Although not identical with the distribution of the proliferation cluster (cluster C) these two clusters show a high degree of similarity.

In Cluster B a high level of expression is seen in Ta grade 3 tumors with frequent recurrences and with Cis but not in the more indolent Ta grade 2 tumors. This cluster contains 11 genes that encode nuclear proteins, such as alpha polymerase, RAD 21, Rb1 and topoisomerase II binding protein.

Cluster C contains genes that are up regulated in both Ta grade 3 with high recurrence rate and CIS, in T2 muscle invasive tumors and in half of the T1 tumors. This cluster show a remarkable tight co-regulation of genes related to cell cycle control and mitosis (Fig.2c). Cyclins, PCNA as well as a number of centromer related proteins are represented in this cluster.

Cluster D holds genes that show a lower than normal expression in muscle invasive stage T2 tumors and Ta grade 3 tumors with Cis, and relatively higher expression in Ta grade 3 and T1 tumors.

Some interesting genes in this cluster are keratin 8 and 19, E-cadherin, Integrin beta 4 and beta 6, and the EGF related genes erb-B2, erb-B3 and EGF receptor pathway substrate 8.

Cluster E holds genes that have a very high expression in Ta grade 2 and 3 without Cis. Among those we find two homeobox proteins A1 and A5, an Insulin like growth factor receptor and Von-Hippel Lindau syndrome protein, as well as an ngi- inducible anti-proliferative protein.

Cluster F shows a tight cluster of genes related to keratinization (Fig. 3). Only two tumor samples (875-1 and 1178-1) show a very high expression of these genes that include keratins 6A, 6B, 14,16,17, small prolin rich proteins 1A and B and 2A and B. A re-evaluation of the pathology slides revealed that only the two samples with high levels of these genes had epidermoid metaplasia. Thus, this cluster of genes explains the gene activation leading to squamous metaplasia as frequently seen by light microscopy in invasive bladder tumors.

Cluster G holds genes that are up-regulated in T2 tumors and have a remarkably consistent high expression level in the Ta grade 3 tumors with Cis that cluster in the invasive branch (Fig. 2g). The cluster is characterized by high levels of genes related to the stroma such as laminin, myosin, caldesmon, collagen, dystrophin, fibronectin, and endoglin. The increased transcription of these genes may indicate a remodeling of the stroma that could reflect signaling from the tumor cells (connective tissue growth factor is included in the cluster) or from infiltrating lymphocytes. It is remarkable that these genes are those that most clearly separate the Ta grade 3 tumors surrounded by Cis from all other Ta grade 3 tumors.

Cluster H is seen as a continuation of cluster G, and like that houses a number of stroma related genes like myosin, tropomyosin, decorin, procollagen and collagens. The prevalence in this cluster of highly expressed genes in both normal biopsies and invasive tumors could indicate that this cluster is reflecting the amount of stroma in the biopsy as that is generally more richly represented in those biopsies.

Cluster I includes genes that are lower in expression in T1 and Ta tumors than in normal urothelium as well as invasive tumors. It contains a large number of genes related to the immune system such as MHC genes, Interleukin receptors, and immunoglobulins. It could be regarded as a measure of the immune response against the tumor, however, the normal biopsies and the muscle invasive tumors look very much alike indicating that it might be a reflection of the amount of stroma in the biopsy. As the level is low in papillomas it cannot be ruled out that papillomas show a reduced immune response for some unknown reason. However, that has to be proven by micro dissection approaches, if that can be done without reducing the RNA quality.

Cluster J includes genes that are highly expressed in invasive tumors, and to some extend in Ta grade 3 with Cis. It houses protease related genes like Matrix metalloproteinase 2 and 9, plasminogen activator urokinase receptor, and urokinase, as well as the cytokine related genes, TNF alpha induced proteins 3 and 6, IL6 and CSF 1, and finally GRO2 and 3 oncogenes. We hypothesize that this cluster is related to the invasive process, however, it is remarkable that the Ta grade 3 tumors with Cis have such a high matrix degrading activity as these tumors have not yet passed the basal membrane. One might suggest that this activity is favoring break down of the basal membrane as well as a fast invasive process when the tumor cells once pass through this. Seen in this light, this cluster may explain why the patients having Cis lesions have such a poor prognosis.

Prediction of bladder tumor stages, generation of a classifier.

An objective class prediction of bladder tumors based on a limited gene-set would be desirable, and could be of potential clinical use. We decided to build a classifier using tumors correctly classified in the three main groups as identified in the cluster dendrogram (Fig. 1a). Consequently, the classifier is based on expression-patterns rather than pathological staging.

We used a maximum likelihood classification method with a cross-validation scheme where one test tumor was removed from the set and a set of predictive genes were selected from the remaining tumor samples for classifying the test tumor. This process was re-iterated for all tumors. Predictive genes that showed the largest possible separation of the three groups were selected for classification, and each tumor was classified according to how close it was to the mean in the three groups (Fig. 4). We classified tumor samples using predictive gene-sets ranging from 10 to 320 genes (Supplementary Information; Table 2). Classification using 80 predictive genes showed the best correlation to pathological staging, more or fewer predictive genes included in the classifier distorted the correlation (Table 1).

Three of the four pTa gr3 tumors with surrounding CIS that clustered as T2+ tumors were classified as T2 and one failed the 5% difference limit (Ta/T1). The solid pT1 tumor (1257-1) that clustered with the muscle invasive tumors was classified as a T1 and the pTa gr3 tumor (1166-1) that clustered with the T1 tumors was classified as a Ta tumor. However, the muscle invasive pT2+ tumor (937-1) previously found in the T1 cluster was also classified as a T1 tumor. This was also the case for tumor 1164-1. It is obvious that the T1 tumors were close to both Ta and T2 tumors, thus forming an intermediate between them (Fig. 4).

Discussion

In this paper we show that applying hierarchical two way clustering to very well characterized clinical specimens can lead to an exact prediction of known and as well as new clinically relevant tumor classes. The specimens were characterized by common pathology features as stage and grade, but also by information on surrounding carcinoma in situ and recurrence pattern through several years. We identified a subset of superficial Ta grade 3 tumors with surrounding Carcinoma *in situ* having properties in common with muscle invasive tumors and indeed clustered together with these. Furthermore, we could distinguish the group of non-recurrent superficial Ta grade2 tumors from Ta grade 3 tumors whit frequent recurrences.

In each class of tumors we identified clusters of genes suggesting some important properties of these classes. For example, we identified a highly increased level of gene transcription factors in Ta grade 3 tumors with frequent recurrences. Three of these transcription factors (TFDP1, TFDP2, and GTF2H4) are involved in cell cycle regulation. In the proliferative cluster that was most prominent in Ta grade 3 with CIS and muscle invasive tumors it was remarkable to observe the many genes related to chromosomal segregation in mitosis. Genes like mitotic kinesin-like protein 1, CDC47, mitotic centromere associated kinesin, centromere protein A, E, and F, and kinesin-like protein 1 all had an up-regulated expression. Whether this is simply reflecting increased cell proliferation, or relates to the well-known aneuploidy found even at early stages of bladder cancer is not known. We do know from a previous study that there are no mutations in the genes related to the anaphase promoting complex, thus a change in expression of genes related to the centromere function offers an alternative explanation that deserves further exploration. These gene products at either RNA or protein level could form important new targets for drug therapy, using for example small molecules that could penetrate the cell wall and exhibit an inhibitory binding to these molecules.

Another important discovery was a cluster of genes related to the stroma and probably indicating stromal remodelling. This cluster was by far most up-regulated in pTa grade 3 tumors with CIS and to almost the same extend in muscle invasive tumors. It contained genes like laminins, hexabrachio, fibulin, myosins, caldesmons, dystrophin, endoglin, collagens IV, V, XV and XVIII, integrins, fibronectin, cadherin, moesin and connective tissue growth factor.

The number of genes used to identify the important clinical classes was originally 1767 but sorting out the genes that were oncology related it could be reduced to only 88 genes. Interestingly, the 88 genes defined three major branches, a Ta, T1 and T2 branch. As with the larger number of genes the T2 branch included the Ta grade3 tumors with CIS. These data points to the fact that it seems possible to classify bladder tumors using a restricted number of genes on a bladder cancer microarray. The smaller number is needed to avoid too much irrelevant noise, and makes interpretation much easier.

Encouraged by this finding we decided to test the strength of suing our gene set as a classifier for bladder cancer samples. Instead of using the pathological staging groups directly we used the three main groups of tumors identified by the cluster analysis. Because of the limited amount of samples in each group we used a cross-validation scheme for classifying the tumors. The obtained classification results showed large similarities to pathological staging when using 80 predictive genes. Furthermore, three of the four Ta gr3 tumors with surrounding CIS, which in the cluster analysis was found close to muscle invasive tumors, were classified as T2 tumors. This is in agreement with the higher risk of disease progression in these patients. In addition, the two muscle invasive tumors (937-1 and 1164-1) classified as T1 tumors were from patients that are still alive after 3 and 2 years respectively.

It will be interesting in the future to follow up on these patients with the aim of evaluating whether the subclasses of T1 and T2 tumors that could be identified hold information on the response to treatment. However it may be more likely that a complete different data set will be needed to generate markers that will predict treatment response.

A commonly observed phenomenon in muscle invasive bladder cancer is squamous metaplasia. Pure squamus cell tumors are relatively rare and have a very poor prognosis with more than 50% of the patients dying within one year [ref]. The two tumors with squamous metaplasia demonstrated clearly some of the genes that are activated in this process, keratins 6A, B 14 , 16, 17 and small proline rich proteins 1A, B and 2B. This corresponds to previous data based on 2-D-gels showing the keratins 6, 14, 16, and 17 highly expressed on the protein level in squamous carcinomas[13]. Furthermore, the small proline rich proteins are present in squamous tissues[14]. Whether the metaplasia is a favorable or unfavorable finding for the disease outcome is not described.

It was interesting that we did not observe systematic alteration in genes related to apoptosis. Reduced apoptosis is supposed to be of major importance in the malignant process as demonstrated in xx cancer by alterations of yy apoptosis related proteins. However, very few apoptosis related genes showed changes in the bladder tumors and none of these in a systematic way. Whether this indicate that apoptosis is of relatively less importance in bladder cancer or that apoptosis is blocked due to inactivating mutations cannot be answered based on the present data. It also emphasizes the fact that we are only registering the level of transcripts by using microarrays. We obtain no information on the quality of the transcripts. These may be harboring inactivating mutations or may be splice variants without biological function; this aspect should always be born in mind when interpreting microarray data.

Previous publications have demonstrated the difference between benign and malignant disease e.g. in the prostate and in the breast. However, this is the first paper to utilize cluster analysis to identify new important classes in a common epithelial carcinoma disease. This was only possible due to the very well characterized clinical material and revitalize the notion that although we have highly sophisticated technologies at hand now, it is still of the utmost importance, and maybe even more important now when thousands of data are obtained from one specimen, that the quality of the specimens to be analyzed is superior.

The very precise class prediction obtained by hierarchical cluster analysis in the present paper is remarkable when taking into account the complete lack of clustering according to stage and grade in clear cell renal carcinomas as recently published[11]. In prostate and colon cancer is was possible to separate benign and malignant diseases[15,16], however, more detailed classification of samples taking into account the disease course and in colon the Dukes stages are yet to come.

We are now able to identify gene clusters that can be used to classify bladder tumors, not only to existing stages and grades but also taking into account surrounding carcinoma in situ and the recurrence pattern. Fabrication of microarrays with the purpose of stratifying patients for specific treatment options is now a possibility.

Methods

Biological material. 40 bladder tumor biopsies were sampled from patients following removal of the necessary amount of tissue for routine pathology examination. The tumors were frozen immediately after surgery and stored at -80°C in a guanidinium thiocyanat solution. All tumors were graded according to Bergkvist *et al.* [17] and re-evaluated by a single pathologist. As normal urothelial reference samples we used a pool of biopsies as well as three single biopsies from patients with prostatic hyperplasia or urinary incontinence. Informed consent was obtained in all cases and protocols were approved by the local scientific ethical committee.

RNA purification and cRNA preparation. Total RNA was isolated from crude tumors biopsies using a Polytron homogenisator and the RNAzol B RNA isolation method (WAK-Chemie Medical GmbH). 10 µg total RNA was used as starting material for the cDNA preparation. The first and second strand cDNA synthesis was performed using the SuperScript Choice System (Life Technologies) according to the manufacturers instructions except using a oligo-dT primer containing a T7 RNA polymerase promoter site. Labeled cRNA was prepared using the BioArray High Yield RNA Transcript Labeling Kit (ENZO). Biotin labeled CTP and UTP (Enzo) were used in the reaction together with unlabeled NTP´s. Following the IVT reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen).

Array hybridization and scanning. 15 µg of cRNA was fragmented at 94^{0}C for 35 min in a fragmentation buffer containing 40 mM Tris-acetate pH 8.1, 100 mM KOAc, 30 mM MgOAc. Prior to hybridization, the fragmented cRNA in a 6xSSPE-T hybridization buffer (1 M NaCl, 10 mM Tris pH 7.6, 0.005% Triton), was heated to 95^{0}C for 5 min and subsequently to 40^{0}C for 5 min before loading onto the Affymetrix probe array cartridge. The probe array was then incubated for 16 h at 45^{0}C at constant rotation (60 rpm). The washing and staining procedure was performed in the Affymetrix Fluidics Station. The probe array was exposed to 10 washes in 6xSSPE-T at 25^{0}C followed by 4 washes in 0.5xSSPE-T at 50^{0}C. The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, final concentration 2 µg/µl (Molecular Probes, Eugene, OR) in 6xSSPE-T for 30 min at 25^{0}C followed by 10 washes in 6xSSPE-T at 25^{0}C. An antibody amplification step was added using normal goat IgG final concentration 0.1 mg/ml (Sigma) and Anti-streptavidin antibody (goat) biotinylated final concentration 3 µg/ml (Vector Laboratories). This was followed by a staining step with a streptavidin-phycoerythrin conjugate, final concentration 2 µg/µl (Molecular Probes, Eugene, OR) in 6xSSPE-T for 30 min at 25^{0}C and 10 washes in 6xSSPE-T at 25^{0}C.

The probe arrays were scanned at 560 nm using a confocal laser-scanning microscope with an argon ion laser as the excitation source (Hewlett Packard GeneArray Scanner G2500A). The readings from the quantitative scanning were analysed by the Affymetrix Gene Expression Analysis Software.

Data analysis. All chips were scaled to a global intensity of 150 units. Expression level ratios between tumors and the normal urothelium reference pool were calculated using the comparison analysis implemented in the Affymetrix GeneChip software. In order to avoid expression ratios based on saturated gene-probes we used the antibody amplified chip-data for genes with an average AvgDiff value below 1000 and the non-amplified data for genes with values equal to or above 1000 in average AvgDiff value. We applied different filtering criteria to the expression data in order to avoid including non-varying and non-measurable genes in the data analysis. First, only genes, which showed significant changes ("Increase" or "Decrease" calls) in expression levels compared to the normal reference pool in at least three samples, were selected Second, only genes with at least three "Present" calls across all experimental samples were selected. Third, we sorted out genes varying less than 2 standard deviations across all samples. The final gene-set contained 1767 genes following filtering. Two-way hierarchical agglomerative cluster analysis was performed using the GeneCluster software[18]. We used average linkage clustering with a modified Pearson correlation as similarity metric. Genes and arrays were median centered and normalized to the magnitude of 1 prior to cluster analysis. The TreeView software was used for visualization of the cluster analysis results[18]. Multidimensional scaling was performed on median centered and normalized data using an implementation in the SPSS statistical software package

Maximum likelihood classifier

We based the classifier on the log-transformed expression level ratios. For these transformed values we used a normal distribution with the mean dependent on the gene and the group (Ta, T1, and T2, respectively) and the variance dependent on the gene only. To classify a sample we calculate the sum over the genes of the squared distance from the sample value to the group mean standardized by the variance. Thus we get a distance to each of the three groups and the sample is classified as belonging to the group where the distance is smallest. When calculating these distances the group means and the variances are estimated from all the samples in the training set excluding the sample being classified. When using a subset of the genes for classification we calculate for each gene the ratio of the variation between the groups to the variation within the groups and select those genes with a high value of this ratio (reference to Dudoit, Fridlyand og Speed). As with any classifier the classifier here can be criticized for being based on a model that is only partly correct. In particular the model does not take into account the correlation among the genes (whether of biological origin or due to artifacts in the data processing). However, some important aspects of the data seems to be captured allowing for a successful classifier.

References

1. Wolf, H., Kakizoe, T., Smith, P.H., et al. *Prog.Clin.Biol.Res.* **221:223-55.**, 223-255 (1986).
2. Cheng, L., Cheville, J.C., Neumann, R.M., et al. *Cancer* **85**, 2469-2474 (1999).
3. Sidransky, D., Frost, P., Von Eschenbach, A., Oyasu, R., Preisinger, A.C. & Vogelstein, B. *N.Engl.J.Med.* **326**, 737-740 (1992).
4. Simon, R., Eltze, E., Schafer, K.L., et al. *Cancer Res.2001.Jan.1.;61.(1.):355.-62.* **61**, 355-362
5. Richter, J., Jiang, F., Gorog, J.P., et al. *Cancer Res.* **57**, 2860-2864 (1997).
6. Sauter, G. & Mihatsch, M.J. *J.Pathol.* **185**, 339-341 (1998).
7. Perou, C.M., Sorlie, T., Eisen, M.B., et al. *Nature 2000.Aug.17.;406.(6797.):747.-52.* **406**, 747-752
8. Alizadeh, A.A., Eisen, M.B., Davis, R.E., et al. *Nature 2000.Feb.3.;403.(6769.):503.-11.* **403**, 503-511
9. Khan, J., Wei, J.S., Ringner, M., et al. *Nat.Med.2001.Jun.;7.(6.):673.-9.* **7**, 673-679
10. Golub, T.R., Slonim, D.K., Tamayo, P., et al. *Science* **286**, 531-537 (1999).
11. Takahashi, M., Rhodes, D.R., Furge, K.A., et al. *Proc.Natl.Acad.Sci.U.S.A.2001.Aug.14.;98.(17.):9754.-9.* **98**, 9754-9759
12. Thykjaer, T., Workman, C., Kruhoffer, M., et al. *Cancer Res.2001.Mar.15.;61.(6.):2492.-9.* **61**, 2492-2499

13. Ostergaard, M., Rasmussen, H.H., Nielsen, H.V., et al. *Cancer Res.* **57,** 4111-4117 (1997).

14. Tesfaigzi, J. & Carlson, D.M. *Cell Biochem.Biophys.* **30,** 243-265 (1999).

15. Alon, U., Barkai, N., Notterman, D.A., et al. *Proc.Natl.Acad.Sci.U.S.A.* **96,** 6745-6750 (1999).

16. Luo, J., Duggan, D.J., Chen, Y., et al. *Cancer Res.2001.Jun.15.;61.(12.):4683.-8.* **61,** 4683-4688

17. Bergkvist, A., Ljungqvist, A. & Moberger, G. *Acta Chir.Scand.* **130,** 371-378 (1965).

18. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc.Natl.Acad.Sci.U.S.A.* **95,** 14863-14868 (1998).

Parallel gene expression monitoring is a powerful tool for the analysis of relations between tumours, for discovering new tumour subgroups, for assigning tumours to pre-defined classes, for identifying co-regulated or tumour stage specific genes, and for predicting outcome[4-17]. In a recent study of bladder cancer, we demonstrated functional groups of genes whose co-regulation formed the basis for separating bladder tumours into superficial and muscle invasive tumours[18]. We now used microarrays with approximately 5000 full-length genes to analyse gene expression and to predict tumour classes in 40 bladder tumours selected from a very large clinical specimen bank holding more than 35.000 samples from bladder cancer patients, prospectively followed for up to six years. The selection was based on the disease course, stage, grade, concomitant carcinoma *in situ* (CIS), and recurrence frequency (number of new tumours per year), in such a way that the selected tumours represent six different groups of patients covering the spectrum from relatively harmless superficial non-recurring papillary Ta grade 2 tumours, to submucosa invasive stage T1 tumours, and finally to primarily muscle invasive T2-4 (T2+) tumours (Table 1; see Supplementary Information Table 1 for the complete disease courses). RNA from tumours and from 4 normal tissue samples (a pool of biopsies from 37 patients and 3 single biopsies) was labelled and hybridised to Affymetrix oligonucleotide microarrays. Scanning identified the expression level of the genes utilising antibody amplification of weakly expressed genes. Genes that did not vary throughout the data-set, e.g. housekeeping genes, were eliminated, and only the 1767 genes (26 %) that showed an expression level change in tumour tissue compared to normal urothelium were subjected to cluster analysis.

A two-way hierarchical cluster analysis of the tumour samples based on the 1767 gene-set remarkably separated all 40 tumours according to conventional pathological stages and grades with only few exceptions (Fig. 1a). We identified two main branches containing the superficial Ta tumours, and the invasive T1 and T2+ tumours. In the superficial branch two sub-clusters of tumours could be identified, one holding 8 tumours that had frequent recurrences and one holding 3 out of the five Ta grade 2 tumours with no recurrences. In the invasive branch, it was notable that four Ta grade 3 tumours clustered tightly with the muscle invasive T2+ tumours. These four Ta tumours, from patients with no previous tumour history, showed concomitant CIS in the surrounding mucosa, indicating that this sub-fraction of Ta tumours has some of the more aggressive features found in muscle invasive tumours. The stage T1 cluster could be separated into three sub-clusters with no clear clinical difference. The one stage T1 grade 3 tumour that clustered with the stage T2+ muscle invasive tumours was the only T1 tumour that showed a solid growth pattern, all others showing papillary growth. Nine out of ten T2+ tumours were found in one single cluster. The remarkable distinct separation of the tumour groups according to stage, with practically no overlap between groups, was also demonstrated by multidimensional scaling analysis (Fig. 1c).

In an attempt to reduce the number of genes needed for class prediction we identified those genes that were scored by the Cancer Genome Anatomy Project (at NCI) as belonging to cancer-related groups such as tumour suppressors, oncogenes, cell cycle, etc. These genes were then selected from the initial 1767 gene-set, and those 88 which showed largest variation (SD of the gene vector >=4), were used for hierarchical clustering of the tumour samples. The obtained clusters was almost identical to the 1767 gene-set cluster dendrogram (Fig. 1b), indicating that the tumour clustering does not simply reflect larger amounts of stromal components in the invasive tumour biopsies.

The clustering of the 1767 genes revealed several characteristic profiles in which there was a distinct difference between the tumour groups (Fig. 1d; black lines identifying clusters a to j). Cluster a, shows a high expression level in all the Ta grade 3 tumours (Fig. 2a) and, as a novel finding, contains genes encoding 8 transcription factors as well as other nuclear genes related to transcriptional activity. Cluster c contains genes that are up-regulated in both Ta grade 3 with high recurrence rate and CIS, in T2+ and some T1 tumours. This cluster shows a remarkable tight co-regulation of gènes related to cell cycle control and mitosis (Fig. 2c). Genes encoding cyclins, PCNA as well as a number of centromere related proteins are present in this cluster. They indicate increased cellular proliferation and may form new targets for small molecule therapy[19]. Cluster f shows a tight cluster of genes related to keratinisation (Fig. 2f). Two tumours (875-1 and 1178-1) had a very high expression of these genes and a re-evaluation of the pathology slides revealed that these were the only two samples to show squamous metaplasia. Thus, activation of this cluster of genes promotes the squamous metaplasia not infrequently seen by light microscopy in invasive bladder tumours. Cluster g contains genes that are up-regulated in T2+ tumours and in the Ta grade 3 tumours with CIS that cluster in the invasive branch (Fig. 2g). This cluster contains genes related to angiogenesis and connective tissue such as laminin, myosin, caldesmon, collagen, dystrophin, fibronectin, and endoglin. The increased transcription of these genes may indicate a profound remodelling of the stroma that could reflect signalling from the tumour cells, from infiltrating lymphocytes, or both. Some of these may also form new drug targets[20]. It is remarkable that these genes are those that most clearly separate the Ta grade 3 tumours surrounded by CIS from all other Ta grade 3 tumours. The presence of adjacent CIS is usually diagnosed by taking a set of eight biopsies from different places in the bladder mucosa. However, the present data clearly indicate that analysis of stroma remodelling genes in the Ta tumours could eliminate this invasive procedure.

The clusters b, d, e, h, i, and j contain genes related to nuclear proteins, cell adhesion, growth factors, stromal proteins, immune system, and proteases, respectively (see Supplementary Information). A summary of the stage related gene expression is shown in Table 2.

An objective class prediction of bladder tumours based on a limited gene-set is clinically usefull. We therefore built a classifier using tumours correctly separated in the three main groups as identified in the cluster dendrogram (Fig. 1a). We used a maximum likelihood classification method with a "leave one out" cross-validation scheme[11 12] in which one test tumour was removed from the set, and a set of predictive genes was selected from the remaining tumour samples for classifying the test tumour. This process was repeated for all tumours. Predictive genes that showed the largest possible separation of the three groups were selected for classification, and each tumour was classified according to how close it was to the mean of the three groups (Fig. 3). The classifier performance was tested using from 1-160 genes in cross-validation loops, and a model using an 80 gene cross-validation scheme showed the best correlation to pathologic staging ($p<10^{-9}$). The 71 genes that were used in at least 75% of the cross validation loops were selected to constitute our final classifier model. To test the class separation performance of the 71 selected genes we compared their performance to those of a permutated set of pseudo-Ta, T1 and T2 tumours. In 500 permutations we only detected two genes with a performance equal to the poorest performing classifying genes (for detailed information on the classifier see Supplementary Information).

The classification using 80 predictive genes in cross-validation loops identified the Ta group with no surrounding CIS and no previous tumor or no previous tumor of a higher stage (Table 1). Interestingly, the Ta tumours surrounded by CIS that were classified as T2 or T1 clearly demonstrate the potential of the classification method for identifying surrounding CIS in a non-invasive way, thereby supplementing clinical and pathologic information.

An objective class prediction of bladder tumours based on a limited gene-set could be of potential clinical use. We therefore built a maximum likelihood classifier using only those tumours (35 out of 40) that showed a group specific expression pattern (Web Figure B). The classifier was evaluated through a "leave one out" cross-validation scheme [11] [12] and predictive genes that showed the largest possible separation of the three groups were selected for classification, and each tumour was classified according to how close it was to the mean of the three groups (Fig. 3*a*). The classifier performance was tested using from 1-200 genes in cross-validation loops, and a model using a 38-gene cross-validation scheme showed the best correlation to pathologic staging (Web Figure C). The 32 genes that were used in at least 75% (27 times) of the cross validations were selected to constitute our final classifier model (Web Table B). Interestingly, some of the Ta tumours surrounded by CIS were classified as T2, thereby supplementing clinical and pathologic information.

We furthermore tested an outcome predictor able to identify the likely presence or absence of recurrence in patients with superficial Ta tumours (see Web Table E for patient disease courses). The optimal number of genes in cross-validation loops was found to be 39 (75% of the samples were correct classified, p<0.006; Web Figure G; Web Table F) and from this we selected those 26 genes (Figure 3*b*) that were used in at least 75% of the cross-validation loops to constitute our final recurrence predictor. Consequently, this set of genes is to be used for predicting recurrence in independent samples. We tested the strength of the predictive genes by permutation analysis (Web Table G).

We present data on expression patterns that classify the benign and muscle-invasive bladder carcinomas. Furthermore, we can identify subgroups of bladder cancer such as Ta tumours with surrounding CIS, Ta tumours with a high probability of progression as well as recurrence, and T2 tumours with squamous metaplasia. As a novel finding, the matrix remodelling gene cluster was specifically expressed in the tumours having the worst prognosis, namely the T2 tumours and tumours surrounded by CIS. For some of these genes new small molecule inhibitors already exist[22], and thus they form drug targets. At present it is not possible clinically to identify patients who will experience recurrence and not recurrenc, but it would be a great benefit to both the patients and the health system by reducing the number of unnecessary control examinations in bladder tumour patients. To determine the optimal gene-set for separating non-recurrent and recurrent tumours, we again applied a cross-validation scheme using from 1-200 genes. We determined the optimal number of genes in cross-validation loops to be 39 (75% of the samples were correct classified, $p<0.01$) and from this we selected those 26 genes (Figure 4) that were used in at least 75% of the cross-validation loops to constitute our final recurrence predictor. Consequently, this set of genes is to be used for predicting recurrence in independent samples. We tested the strength of the predictive genes by performing 500 permutations of the arrays. This revealed that for most of our predictive genes we would only in a small number of the new pseudo-groups obtain at least as good predictors as in the real groups (see further details in Supplementary Information).

We present data on expression patterns that classify the different well-known clinical stages of bladder carcinoma. Furthermore, we can classify subgroups of bladder cancers such as Ta tumours with surrounding CIS, Ta tumours with recurrence potential, and T2 tumours with squamous metaplasia. This has implications for epithelial cancers in general as these may be subdivided into a larger number of subclasses than has previously been expected, due to the sensitive way in which microarrays detect even minor tumour variations. As a novel finding, the matrix remodelling gene cluster was specifically expressed in the tumours having the worst prognosis, namely the T2 tumours and tumours surrounded by CIS. Furthermore, another novel distinct molecular feature was the high expression of transcription related genes in Ta tumours.

The ability to classify bladder tumours, to identify Ta tumours that will recur and to make a non-invasive diagnosis of CIS in the bladder is of immediate clinical relevance. In a larger perspective many of the differentially expressed genes form new drug targets, e.g. the matrix remodelling related genes, for some of which new small molecule inhibitors already exist[22].

Methods

Biological material. 66 bladder tumour biopsies were sampled from patients following removal of the necessary amount of tissue for routine pathology examination. The tumours were frozen immediately after surgery and stored at -80°C in a guanidinium thiocyanate solution. All tumours were graded according to Bergkvist *et al.*[23] and re-evaluated by a single pathologist. As normal urothelial reference samples we used a pool of biopsies (from 37 patients) as well as three single bladder biopsies from patients with prostatic hyperplasia or urinary incontinence. Informed consent was obtained in all cases and protocols were approved by the local scientific ethical committee.

cRNA preparation, GeneChip hybridisation and scanning. Target cRNAs were synthesised and hybridised to Affymetrix GeneChip Hu6800 oligonucleotide microarrays as recommended. See Supplementary Information for detailed descriptions.

Class discovery using hierarchical clustering. All microarray results were scaled to a global intensity of 150 units using the Affymetrix GeneChip software. Other ways of array normalisation exist[24], however, using the dCHIP approach did not change the expression profiles of the obtained classifier genes in this study (results not shown). For hierarchical cluster analysis and molecular classification procedures we used expression level ratios between tumours and the normal urothelium reference pool calculated using the comparison analysis implemented in the Affymetrix

GeneChip software. In order to avoid expression ratios based on saturated gene-probes, we used the antibody amplified expression-data for genes with a mean Average Difference value across all samples below 1000 and the non-amplified expression-data for genes with values equal to or above 1000 in mean Average Difference value across all samples. Consequently, gene expression levels across all samples were either from the amplified or the non-amplified expression-data. We applied different filtering criteria to the expression data in order to avoid including non-varying and very low expressed genes in the data analysis. Firstly, we selected only genes that showed significant changes in expression levels compared to the normal reference pool in at least three samples. Secondly, only genes with at least three "Present" calls across all samples were selected. Thirdly, we eliminated genes varying less than 2 standard deviations across all samples. The final gene-set contained 1767 genes following filtering. Two-way hierarchical agglomerative cluster analysis was performed using the Cluster software[25]. We used average linkage clustering with a modified Pearson correlation as similarity metric. Genes and arrays were median centred and normalised to the magnitude of 1 prior to cluster analysis. The TreeView software was used for visualisation of the cluster analysis results[25]. Multidimensional scaling was performed on median centred and normalised data using an implementation in the SPSS statistical software package.

Tumour stage classifier. We based the classifier on the log-transformed expression level ratios. For these transformed values we used a normal distribution with the mean dependent on the gene and the group (Ta, T1, and T2, respectively) and the variance dependent on the gene only. For each gene we calculated the ratio of the variation between the groups to the variation within the groups, and selected those genes with a high ratio value. To classify a sample, we calculated the sum over the genes of the squared distance from the sample value to the group mean, standardised by the variance. Thus, we got a distance to each of the three groups and the sample was classified as belonging to the group in which the distance was smallest. When calculating these distances the group means and the variances were estimated from all the samples in the training set excluding the sample being classified.

Recurrence prediction using a supervised learning method. Average Difference values were generated using the Affymetrix GeneChip software and all values below 20 were set to 20 to avoid very low and negative numbers. We only included genes that had a "Present" call in at least 7 samples and genes that showed intensity variation (Max-Min>100, Max/Min>2). The values were log transformed and rescaled. We used a supervised learning method essentially as described[11]. Genes were selected using t-test statistics and cross-validation and sample classification was performed as described above.

Immunohistochemistry. Tumour tissue microarrays were prepared essentially as described[26], with four representative 0.6 mm paraffin cores from each study case. Immunohistochemical staining was performed using standard highly sensitive techniques after appropriate heat-induced antigen retrieval. Primary polyclonal goat antibodies against Smad 6 (S-20) and cyclin G2 (N-19) were from Santa Cruz Biotechnology, Santa Cruz, CA. Antibodies to p53 (monoclonal DO-7) and Her-2 (polyclonal anti-c-erbB-2) were from Dako A/S, Glostrup, Denmark. Ki-67 monoclonal antibody (MIBI) was from Novocastra Laboratories Ltd, Newcastle-upon-Tyne, UK.

Methods

Biological material. 66 bladder tumour biopsies were sampled from patients following removal of the necessary amount of tissue for routine pathology examination. The tumours were frozen immediately after surgery and stored at -80°C in a guanidinium thiocyanate solution. All tumours were graded according to Bergkvist *et al.*[23] and re-evaluated by a single pathologist. As normal urothelial reference samples we used a pool of biopsies (from 37 patients) as well as three single bladder biopsies from patients with prostatic hyperplasia or urinary incontinence. Informed consent was obtained in all cases and protocols were approved by the local scientific ethical committee.

RNA purification and cRNA preparation. Total RNA was isolated from crude tumour biopsies using a Polytron homogenisator and the RNAzol B RNA isolation method (WAK-Chemie Medical

GmbH). 10 µg total RNA was used as starting material for the cDNA preparation. The first and second strand cDNA synthesis was performed using the SuperScript Choice System (Life Technologies) according to the manufacturers instructions except using an oligo-dT primer containing a T7 RNA polymerase promoter site. Labelled cRNA was prepared using thė BioArray High Yield RNA Transcript Labelling Kit (Enzo). Biotin labelled CTP and UTP (Enzo) were used in the reaction together with unlabeled NTP's. Following the IVT reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen).

Array hybridisation and scanning. 15 µg of cRNA was fragmented at 94^{0}C for 35 min in a fragmentation buffer containing 40 mM Tris-acetate pH 8.1, 100 mM KOAc, 30 mM MgOAc. Prior to hybridisation, the fragmented cRNA in a 6xSSPE-T hybridisation buffer (1 M NaCl, 10 mM Tris pH 7.6, 0.005% Triton), was heated to 95^{0}C for 5 min and subsequently to 45^{0}C for 5 min before loading onto the Affymetrix probe array cartridge (HuGeneFL). The probe array was then incubated for 16 h at 45^{0}C at constant rotation (60 rpm). The washing and staining procedure was performed in the Affymetrix Fluidics Station. The probe array was exposed to 10 washes in 6xSSPE-T at 25^{0}C followed by 4 washes in 0.5xSSPE-T at 50^{0}C. The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, final concentration 2 µg/µl (Molecular Probes, Eugene, OR) in 6xSSPE-T for 30 min at 25^{0}C followed by 10 washes in 6xSSPE-T at 25^{0}C. The probe arrays were scanned at 560 nm using a confocal laser-scanning microscope (Hewlett Packard GeneArray Scanner G2500A). The readings from the quantitative scanning were analysed by the Affymetrix Gene Expression Analysis Software. An antibody amplification step followed using normal goat IgG as blocking reagent, final concentration 0.1 mg/ml (Sigma) and biotinylated anti-streptavidin antibody (goat), final concentration 3 µg/ml (Vector Laboratories). This was followed by a staining step with a streptavidin-phycoerythrin conjugate, final concentration 2 µg/µl (Molecular Probes, Eugene, OR) in 6xSSPE-T for 30 min at 25^{0}C and 10 washes in 6xSSPE-T at 25^{0}C. The arrays were then subjected to a second scan under similar conditions as described above.

Class discovery using hierarchical clustering. All microarray results were scaled to a global intensity of 150 units using the Affymetrix GeneChip software. Other ways of array normalisation exist[24], however, using the dCHIP approach did not change the expression profiles of the obtained classifier genes in this study (results not shown). For hierarchical cluster analysis and molecular classification procedures we used expression level ratios between tumours and the normal urothelium reference pool calculated using the comparison analysis implemented in the Affymetrix GeneChip software. In order to avoid expression ratios based on saturated gene-probes, we used the antibody amplified expression-data for genes with a mean Average Difference value across all samples below 1000 and the non-amplified expression-data for genes with values equal to or above 1000 in mean Average Difference value across all samples. Consequently, gene expression levels across all samples were either from the amplified or the non-amplified expression-data. We applied different filtering criteria to the expression data in order to avoid including non-varying and very low expressed genes in the data analysis. Firstly, we selected only genes that showed significant changes in expression levels compared to the normal reference pool in at least three samples. Secondly, only genes with at least three "Present" calls across all samples were selected. Thirdly, we eliminated genes varying less than 2 standard deviations across all samples. The final gene-set contained 1767 genes following filtering. Two-way hierarchical agglomerative cluster analysis was performed using the Cluster software[25]. We used average linkage clustering with a modified Pearson correlation as similarity metric. Genes and arrays were median centred and normalised to the magnitude of 1 prior to cluster analysis. The TreeView software was used for visualisation of the cluster analysis results[25]. Multidimensional scaling was performed on median centred and normalised data using an implementation in the SPSS statistical software package.

Tumour stage classifier. We based the classifier on the log-transformed expression level ratios. For these transformed values we used a normal distribution with the mean dependent on the gene and the group (Ta, T1, and T2, respectively) and the variance dependent on the gene only. For each gene we calculated the variation within the groups (W) and the three variations between two groups

(B(Ta/T1), B(Ta/T2), B(T1/T2)) and used the three ratios B/W to select genes. We selected those genes having a high value of B(Ta/T1)/W, those genes having a high value of B(Ta/T2)/W, and those genes with a high value of B(T1/T2)/W. To classify a sample, we calculated the sum over the genes of the squared distance from the sample value to the group mean, standardised by the variance. Thus, we got a distance to each of the three groups and the sample was classified as belonging to the group in which the distance was smallest. When calculating these distances the group means and the variances were estimated from all the samples in the training set excluding the sample being classified.

Validation of the tumour stage classifier. The performance of the classifier was validated using another set of bladder tumour expression data obtained from customised oligonucleotide Affymetrix GeneChips carrying PM probes only. First, we translated all accession numbers on both oligonucleotide microarrays into UG-clusters and selected those gene-probes present on both arrays (4416 probe-sets). To make comparisons between the two microarray types we used only the PM probe values from the original data set. We rescaled all the log (average PM) values and used the pool of normal bladder biopsies from 37 patients, which were analyses on both array platforms, to calculate log fold-change expression values. We recalculated the group means and the variances for each gene used in the classifier and based the classification on 29 genes from the optimal classifier in the cross-validation step for the original dataset. For the new samples the distances to each of the three groups was calculated and the sample was classified as belonging to the group for which the distance was smallest.

Recurrence prediction using a supervised learning method. Average Difference values were generated using the Affymetrix GeneChip software and all values below 20 were set to 20 to avoid very low and negative numbers. We only included genes that had a "Present" call in at least 7 samples and genes that showed intensity variation (Max-Min>100, Max/Min>2). The values were log transformed and rescaled. We used a supervised learning method essentially as described[11].

Genes were selected using t-test statistics and cross-validation and sample classification was performed as described above.

Immunohistochemistry. Tumour tissue microarrays were prepared essentially as described[26], with four representative 0.6 mm paraffin cores from each study case. Immunohistochemical staining was performed using standard highly sensitive techniques after appropriate heat-induced antigen retrieval. Primary polyclonal goat antibodies against Smad 6 (S-20) and cyclin G2 (N-19) were from Santa Cruz Biotechnology. Antibodies to p53 (monoclonal DO-7) and Her-2 (polyclonal anti-c-erbB-2) were from Dako A/S. Ki-67 monoclonal antibody (MIBI) was from Novocastra Laboratories Ltd. Staining intensity was scored at four levels, Negative, Weak, Moderate and Strong by an experienced pathologist who considered both colour intensity and number of stained cells, and who was unaware of array results.

References

1. Pisani, P., Parkin, D. M., Bray, F., & Ferlay, J. Estimates of the worldwide mortality from 25 cancers in 1990. *Int.J.Cancer* **83**, 18-29 (1999).

2. Wolf, H. *et al.* Bladder tumors. Treated natural history. *Prog.Clin.Biol.Res.* **221:223-55.**, 223-255 (1986).

3. Cheng, L. *et al.* Survival of patients with carcinoma in situ of the urinary bladder. *Cancer* **85**, 2469-2474 (1999).

4. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

5. Khan, J. *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat.Med.* **7**, 673-679 (2001).

6. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).

7. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).

8. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat.Genet.* **24**, 227-235 (2000).

9. Takahashi, M. *et al.* Gene expression profiling of clear cell renal cell carcinoma: Gene identification and prognostic classification. *Proc.Natl.Acad.Sci.U.S.A.* **98**, 9754-9759 (2001).

10. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc.Natl.Acad.Sci.U.S.A* **98**, 10869-10874 (2001).

11. Shipp, M. A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat.Med.* **8** , 68-74 (2002).

12. 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).

13. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc.Natl.Acad.Sci.U.S.A.* **96**, 6745-6750 (1999).

14. Luo, J. *et al.* Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.* **61**, 4683-4688 (2001).

15. Notterman, D. A., Alon, U., Sierk, A. J., & Levine, A. J. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* **61**, 3124-3130 (2001).

16. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc.Natl.Acad.Sci.U.S.A.* **96**, 9212-9217 (1999).

17. Dhanasekaran, S. M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822-826 (2001).

18. Thykjaer, T. *et al.* Identification of gene expression patterns in superficial and invasive human bladder cancer. *Cancer Res.* **61**, 2492-2499 (2001).

19. Seymour, L. Novel anti-cancer agents in development: exciting prospects and new challenges. *Cancer Treat.Rev.* **25**, 301-312 (1999).

20. Fox, S. B., Gasparini, G., & Harris, A. L. Angiogenesis: pathological, prognostic, and growth-factor pathways and their link to trial design and anticancer drugs. *Lancet Oncol.* **2**, 278-289 (2001).

21. Ørntoft, T. F. *et al.* Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. *Mol Cell Proteomics.* **1**, 37-45 (2002).

22. Kerr, J. S., Slee, A. M., & Mousa, S. A. Small molecule alpha(v) integrin antagonists: novel anticancer agents. *Expert.Opin.Investig.Drugs* **9**, 1271-1279 (2000).

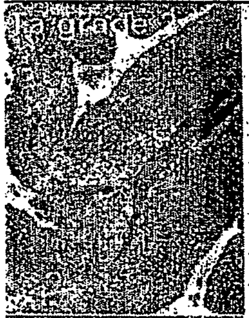
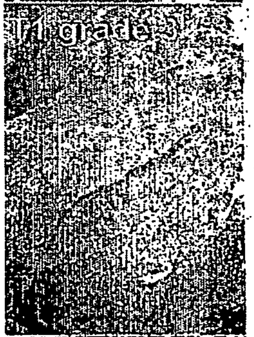
23. Bergkvist, A., Ljungqvist, A., & Moberger, G. Classification of bladder tumours based on the cellular pattern. Preliminary report of a clinical-pathological study of 300 cases with a minimum follow-up of eight years. *Acta Chir.Scand.* **130**, 371-378 (1965).

24. Li, C. & Hung, W. W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, RESEARCH0032 (2001).

25. Eisen, M. B., Spellm[illegible]. T., Brown, P. O., & Botstein, D. Clust[illegible]lysis and display of genome-wide expression patterns. *Proc.Natl.Acad.Sci.U.S.A.* 95, 14863-14868 (1998).

26. Kononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat.Med.* 4, 844-847 (1998).

Table 1 • Clinical data on disease courses and results of molecular classification

Tumours[a]	Patient	Previous tumours	Tumour analysed	Subsequent tumours	Carcinoma *in situ*[b]	Reviewed histology[c]	Molecular classifier[d]		
							320	80	20
	Ta grade II tumours – no progression								
	709-1		Ta gr2		No	Ta gr3	Ta	Ta	Ta
	968-1		Ta gr2	1 Ta	No		Ta/T1	Ta	Ta
	934-1		Ta gr2		No		T1	Ta	Ta
	928-1		Ta gr2		No		Ta	Ta	T1
	930-1		Ta gr2		No		Ta	Ta	Ta
	Ta grade III tumours – no prior T1 tumour or CIS								
	989-1		Ta gr3		No		Ta	Ta	Ta
	1264-1		Ta gr3	3 Ta	No		Ta	Ta	Ta
	876-5	4 Ta	Ta gr3		No		Ta	Ta	Ta
	669-7	5 Ta	Ta gr3	4 Ta	No	Ta gr2	Ta	Ta	Ta
	716-2	1 Ta	Ta gr3	2 Ta	No		Ta	Ta	Ta
	Ta grade III tumours – no prior T1 tumour but CIS in selected site biopsies								
	1070-1		Ta gr3	1 Ta	Subsequent visit		Ta	Ta	Ta
	956-2		Ta gr3	1 Ta	Sampling visit		T2	T2	T2/T1
	1062-2		Ta gr3	1 T1	Sampling visit		T2/Ta	T1/Ta	Ta
	1166-1		Ta gr3		Sampling visit		Ta/T1	Ta	Ta
	1330-1		Ta gr3		Sampling visit		T2	T2	Ta
	Ta grade III tumours – a prior T1 tumour and CIS in selected site biopsies								
	747-7	5 Ta, 1 T1	Ta gr3	3 Ta	Sampling visit		Ta	Ta	Ta
	112-10	7 Ta, 2 T1	Ta gr3	2 Ta, 4 T1	Previous visit		Ta	Ta	Ta
	320-7	1 Ta, 2 T1	Ta gr3	2 Ta	Sampling visit		T2	T2	Ta
	967-3	2 T1	Ta gr3	1 T1	Sampling visit		Ta	Ta	Ta
	T1 grade III tumours – no prior muscle invasive tumour								
	625-1		T1 gr3		No		T1	T1	T1
	847-1		T1 gr3		No		T1	T1	T1
	1257-1		T1 gr3		Sampling visit		T1	T1	T1
	919-1		T1 gr3		No		T1	T1	T1
	880-1		T1 gr3	4 Ta	No		T1	T1	T1
	812-1		T1 gr3		No		T1	T1	T1
	1269-1		T1 gr3		No	No review	T1	T1	T1
	1083-2	1 Ta	T1 gr3		No	No review	T1	T1	T1
	1238-1		T1 gr3	1 Ta, 1 T2+	No		T1	T1	T1
	1065-1		T1 gr3		Subsequent visit	No review	T1	T1	T1
	1134-1		T1 gr3	3 T1	Sampling visit	T2 gr3	T1	T1	T1
	T2+ grade III/IV tumours – only primary tumours								
	1164-1		T2+ gr4		No	T2+ gr3	T2/T1	T1	T1
	1032-1		T2+ gr?		ND	No review	T2	T2	T2
	1117-1		T2+ gr3		ND		T2	T2	T1
	1178-1		T2+ gr3		ND		T2	T2	T2

1078-1	T2+ gr3		ND		T2	T2	T2
875-1	T2+ gr3		No		T2	T2	T2
1044-1	T2+ gr3	1 T2+	ND		T2	T2	T2
1133-1	T2+ gr3		ND		T2	T2	T2
1068-1	T2+ gr3		No		T2	T2	T2
937-1	T2+ gr3		ND	No review	T1	T1	T1

[a] Examples of tumour histology.

[b] Carcinoma *in situ* detected in selected site biopsies at the time of sampling tumour tissue for the arrays or at previous or subsequent visits.

[c] All tumours were reviewed by a single uro-pathologist and any change compared to the routine classification is listed.

[d] Molecular classification based on 320, 80, and 20 genes cross-validation loops.

Table 2 • Summary of stage related gene expression

Functional gene clusters[a]

Tumour stage	Transcription	Nuclear processes	Proliferation	Matrix remodelling	Extracellular matrix	Immune system
Ta gr2	↑	-	-	-	↓↓	↓
Ta gr3	↑↑↑	↑↑	↑↑	-	↓↓	↓
T1 gr3	↑[b]	-	↑↑[b]	-	↓	↑[b]
T2 gr3	↑	-	↑↑↑	↑↑↑	↑	↑
Ta gr3 + CIS	↑↑↑	↑↑	↑↑↑	↑↑↑	↑	↑

a For a detailed description of gene clusters see Supplementary Information page 6.

b An increase in gene expression was only found in about half of the samples analysed.

Figure legends

Fig. 1 Two-way hierarchical clustering and multidimensional scaling analysis of gene expression data from 40 bladder tumour biopsies. **a,** Tumour cluster dendrogram based on the 1767 gene-set. CIS annotations following the sample names indicate concomitant carcinoma in situ. Tumour recurrence rates are shown to the right of the dendrogram as + and ++ indicating moderate and high recurrence rates, respectively, while no sign indicates no or moderate recurrence. **b,** Tumour cluster dendrogram based on 88 cancer related genes. **c,** 2D plot of multidimensional scaling analysis of the 40 tumours based on the 1767 gene-set. The colour code identifies the tumour samples from the cluster dendrogram (Fig. 1a). **d,** Two-way cluster analysis diagram of the 1767 gene-set. Each row in the diagram represents a gene and each column a tumour sample. The colour saturation represents differences in gene expression across the tumour samples; yellow indicates higher expression of the gene compared to the median expression (black) and blue indicates lower expression of the gene compared to the median expression. The colour intensities indicate degrees of gene-regulation. The sidebars to the right of the diagram represent gene clusters **a-j** and normal 1-3 in the left side indicate the three normal biopsies and normal 4 indicates the pool of biopsies from 37 patients.

Fig. 2 Enlarged view of the gene clusters **a, c, f,** and **g.** The dendrogram at the top is identical to Fig. 1a. **a,** Cluster of transcription factors and other nuclear associated genes. **c,** Cluster of genes involved in proliferation and cell cycle control. **f,** Gene expression pattern and corresponding area with squamous metaplasia in urothelial carcinoma. The yellow colour indicates genes up-regulated in samples 1178-1 and 875-1, the only two samples with squamous cell metaplasia. **g,** Cluster of genes involved in angiogenesis and matrix remodelling.

Fig. 3 Molecular classification of tumour samples using 80 predictive genes in each cross-validation loop. Each classification is based on the closeness to the mean in the three classes. Samples marked

with * were not used to build the classifier. The scale indicates the distance from the samples to the classes in the classifier, measured in weighted squared Euclidean distance.

Fig. 4 Gene expression patterns of the 26 genes that we found to be optimal for prediction of superficial tumour recurrence. The best predictors of recurrence are listed at the top and bottom of the diagram. For each gene the number of times it was used in the 31 cross-validation loops is listed to the right together with the unigene-cluster number (see more details in Supplementary Information).

Supplementary Information

Identifying distinct classes of bladder carcinoma using microarrays.

Lars Dyrskjøt Andersen, Thomas Thykjaer, Mogens Kruhøffer, Jens Ledet Jensen, Niels Marcussen, Stephen Hamilton-Dutoit, Hans Wolf & Torben F. Ømtoft

Contents:

Methods

The following paragraphs contain supplementary information about cRNA preparation, chip hybridisation and scanning protocols not described in the paper.

RNA purification and cRNA preparation

Total RNA was isolated from crude tumour biopsies using a Polytron homogenisator and the RNAzol B RNA isolation method (WAK-Chemie Medical GmbH). 10 μg total RNA was used as starting material for the cDNA preparation. The first and second strand cDNA synthesis was performed using the SuperScript Choice System (Life Technologies) according to the manufacturers instructions except using an oligo-dT primer containing a T7 RNA polymerase promoter site. Labelled cRNA was prepared using the BioArray High Yield RNA Transcript Labelling Kit (Enzo). Biotin labelled CTP and UTP (Enzo) were used in the reaction together with unlabeled NTP's. Following the IVT reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen).

Array hybridisation and scanning

15 μg of cRNA was fragmented at 94^{0}C for 35 min in a fragmentation buffer containing 40 mM Tris-acetate pH 8.1, 100 mM KOAc, 30 mM MgOAc. Prior to hybridisation, the fragmented cRNA in a 6xSSPE-T hybridisation buffer (1 M NaCl, 10 mM Tris pH 7.6, 0.005% Triton), was heated to 95^{0}C for 5 min and subsequently to 45^{0}C for 5 min before loading onto the Affymetrix probe array cartridge. The probe array was then incubated for 16 h at 45^{0}C at constant rotation (60 rpm). The washing and staining procedure was performed in the Affymetrix Fluidics Station. The probe array was exposed to 10 washes in 6xSSPE-T at 25^{0}C followed by 4 washes in 0.5xSSPE-T at 50^{0}C. The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, final concentration 2 μg/μl (Molecular Probes, Eugene, OR) in 6xSSPE-T for 30 min at 25^{0}C followed by 10 washes in 6xSSPE-T at 25^{0}C. The probe arrays were scanned at 560 nm using a confocal laser-scanning microscope (Hewlett Packard GeneArray Scanner G2500A). The readings from the quantitative scanning were analysed by the Affymetrix Gene Expression Analysis Software. An antibody amplification step followed using normal goat IgG as blocking reagent, final concentration 0.1 mg/ml (Sigma) and biotinylated anti-streptavidin antibody (goat), final concentration 3 μg/ml (Vector Laboratories). This was followed by a staining step with a streptavidin-phycoerythrin conjugate, final concentration 2 μg/μl (Molecular Probes, Eugene, OR) in 6xSSPE-T for 30 min at 25^{0}C and 10 washes in 6xSSPE-T at 25^{0}C. The arrays were then subjected to a second scan under similar conditions as described above.

Samples

This part contains information about all the samples used for expression profiling. All samples used were obtained fresh from surgery and the tumour material for expression profiling was frozen immediately at -80°C after removing material for histopathological analysis. As reference we used biopsies from normal urothelium from donors with prostatic hyperplasia or incontinence.

Patient disease course information – class discovery

We selected tumours from the entire spectrum of bladder carcinoma for expression profiling in order to discover the molecular classes of the disease. The tumours analysed are listed in Table 1 below together with the available patient disease course information.

Table 1. Disease course information of all patients involved.

Group	Patient	Previous tumours	Tumour examined on array	Pattern	Reviewed histology	Subsequent tumours	Carcinoma *in situ**
A	709-1		Ta gr 2 (200297)	Papillary	Ta gr3		no
	968-1		Ta gr 2 (011098)	Papillary	+	Ta gr 2 (150101)	no
	934-1		Ta gr 2 (220798)	Papillary	+		no
	928-1		Ta gr 2 (240698)	Papillary	+		no
	930-1		Ta gr 2 (300698)	Papillary	+		no
B	989-1		Ta gr 3 (281098)	Papillary	+		no
	1264-1		Ta gr 3 (130600)	Papillary	+	Ta gr 2 (231000) Ta gr 2 (220101) Ta gr 2 (300401)	no
	876-5	Ta gr 2 (230398) Ta gr 2 (271098) Ta gr 2 (090699) Ta gr 2 (011199)	Ta gr 3 (170400)	Papillary	+		no
	669-7	Ta gr 2 (101296) Ta gr 2 (150897) Ta gr 1 (161297) Ta gr 3 (270498) Ta gr 2 (220299)	Ta gr 3 (230899)	Papillary	Ta gr2	Ta gr 2 (120100) Ta gr 2 (250500) Ta gr 2 (250900) Ta gr 2 (050201)	no
	716-2	Ta gr 2 (070397)	Ta gr 3 (230497)	Papillary	+	Ta gr 2 (040697) Ta gr 1 (170698)	no
C	1070-1		Ta gr 3 (150399)	Papillary	+	Ta gr 3 (291099)	Subsequent visit
	956-2		Ta gr 3 (061299)	Papillary	+	Ta gr 3 (061200)	Sampling visit
	1062-2		Ta gr 3 (120799)	Papillary	+	T1 gr 3 (161199)	Sampling visit
	1166-1		Ta gr 3 (271099)	Papillary	+		Sampling visit
	1330-1		Ta gr 3 (311000)	Papillary	+		Sampling visit
D	112-10	Ta gr 2 (070794) Ta gr 3 (011294) T1 gr 3(150695) Ta gr 3 (121095) T1 gr 3(040396) Ta gr 2 (200896) Ta gr 2 (111296) Ta gr 2 (230497) Ta gr 2 (030997)	Ta gr 3 (060198)	Papillary	+	Ta gr 3 (110698) T1 gr 3 (191098) Ta gr 3 (240299) T1 gr 3 (050799) T1 gr 3 (081199) T1 gr 3 (180400)	Previous visit
	320-7	T1 gr 3 (011194) T1 gr 3 (150896) Ta gr 3 (100897)	Ta gr 3 (290997)	Papillary	+	Ta gr 3 (290198) Ta gr 3 (290698)	Sampling visit
	747-7	Ta gr 2 (010597) Ta gr 2 (220597) Ta gr 2 (230997) Ta gr 2 (260198) T1 gr 3 (270498) Ta gr 2 (170898)	Ta gr 3 (161298)	Papillary	+	Ta gr 2 (050599) Ta gr 2 (280999) Ta gr 2,(141299)	Sampling visit
	967-3	T1 gr 3 (280998) T1 gr 3 (250199)	Ta gr 3 (140699)	Papillary	+	T1 gr 3 (080999)	Sampling visit
E	625-1		T1 gr 3 (200996)	Papillary	+		No
	847-1		T1 gr 3 (210198)	Papillary	+		No

	1257-1		T1 gr 3 (240500)	Solid	+		Sampling visit
	919-1		T1 gr 3 (220698)	Papillary	+		No
	880-1		T1 gr 3 (300398)	Papillary	+	Ta gr 2 (091198) Ta gr 1 (090399) Ta gr 2 (050900) Ta gr 2 (190301)	No
	812-1		T1 gr 3 (061098)	Papillary	+		No
	1269-1		T1 gr 3 (230600)	Papillary	-		No
	1083-2	Ta gr 2 (280499)	T1 gr 3 (120599)	Papillary	-		No
	1238-1		T1 gr 3 (020500)	Papillary	+	T2 gr 3 (211100) Ta gr 2 (211100)	No
	1065-1		T1 gr 3 (160399)	Papillary	-		Subsequent visit
	1134-1		T1 gr 3 (181099)	Papillary	T2 gr3	T1 gr 3 (280200) T1 gr 3 (020500) T1 gr 3 (131100)	Sampling visit
F	1164-1		T2+ gr 4 (101299)	Solid	gr 3		No
	1032-1		T2+ gr ? (050199)	Mixed	-		Not measured
	1117-1		T2+ gr 3 (010999)	Solid	+		Sampling visit
	1178-1		T2+ gr 3 (200100)	Solid	+		Not measured
	1078-1		T2+ gr 3 (120499)	Solid	+		Not measured
	875-1		T2+ gr 3 (180398)	Solid	+		No
	1044-1		T2+ gr 3 (010299)	Solid	+	T2+ gr 3 (060999)	Not measured
	1133-1		T2+ gr 3 (081099)	Solid	+		Not measured
	1068-1		T2+ gr 3 (220399)	Solid	+		No
	937-1		T2+ gr 3 (280798)	Solid	-		Not measured

Group A: Ta gr2 tumours – no recurrence within 2 years.

Group B: Ta gr3 tumours – no prior T1 tumour and no carcinoma *in situ* in random biopsies.

Group C: Ta gr3 tumours – no prior T1 tumour but carcinoma *in situ* in random biopsies. Group D: Ta gr3 tumours – a prior T1 tumour and carcinoma *in situ* in random biopsies. Group E: T1 gr3 tumours – no prior T2+ tumour. Group F: T2+ tumours gr3/4 – only primary tumours.

* Carcinoma *in situ* detected in selected site biopsies at previous, sampling or subsequent visits.

Patient disease course information – recurrence vs. no recurrence

From the hierarchical cluster analysis of the tumour samples we found that the tumours with a high recurrence frequency were separated from the tumours with low recurrence frequency. To study this further we profiled two groups of Ta tumours- 15 tumours with low recurrence frequency and 16 tumours with high recurrence frequency. To avoid influence from other tumour characteristics we only used tumours that showed the same growth pattern and tumours that showed no sign of concomitant carcinoma *in situ*. Furthermore, the tumours were all primary tumours. The tumours used for identifying genes differentially expressed in recurrent and non-recurrent tumours are listed in Table 2 below.

Table 2 Disease course information of all patients involved.

Group	Patient	Tumour (date)	Pattern	Carcinoma *in situ*	Time to recurrence
A	968-1	Ta gr2	Papillary	no	27 month
A	928-1	Ta gr2	Papillary	no	38 month.
A	934-1	Ta gr2 (220798)	Papillary	no	-
A	709-1	Ta gr2 (210798)	Papillary	no	-
A	930-1	Ta gr2 (300698)	Papillary	no	-
A	524-1	Ta gr2 (201095)	Papillary	no	-
A	455-1	Ta gr2 (060695)	Papillary	no	-
A	370-1	Ta gr2 (100195)	Papillary	no	-
A	810-1	Ta gr2 (031097)	Papillary	no	-
A	1146-1	Ta gr2 (231199)	Papillary	no	-
A	1161-1	Ta gr2 (101299)	Mixed	no	-
A	1006-1	Ta gr2 (231198)	Papillary	no	-
A	942-1	Ta gr2	Papillary	no	24 month.
A	1060-1	Ta gr2	Papillary	no	36 month.
A	1255-1	Ta gr2	Papillary	no	24 month.
B	441-1	Ta gr2	Papillary	no	6 month.
B	780-1	Ta gr2	Papillary	no	2 month.
B	815-2	Ta gr2	Papillary	no	6 month.
B	829-1	Ta gr2	Papillary	no	4 month.
B	861-1	Ta gr2	Papillary	no	4 month.
B	925-1	Ta gr2	Papillary	no	5 month.
B	1008-1	Ta gr2	Papillary	no	5 month.
B	1086-1	Ta gr2	Papillary	no	6 month.
B	1105-1	Ta gr2	Papillary	no	8 month.
B	1145-1	Ta gr2	Papillary	no	4 month.

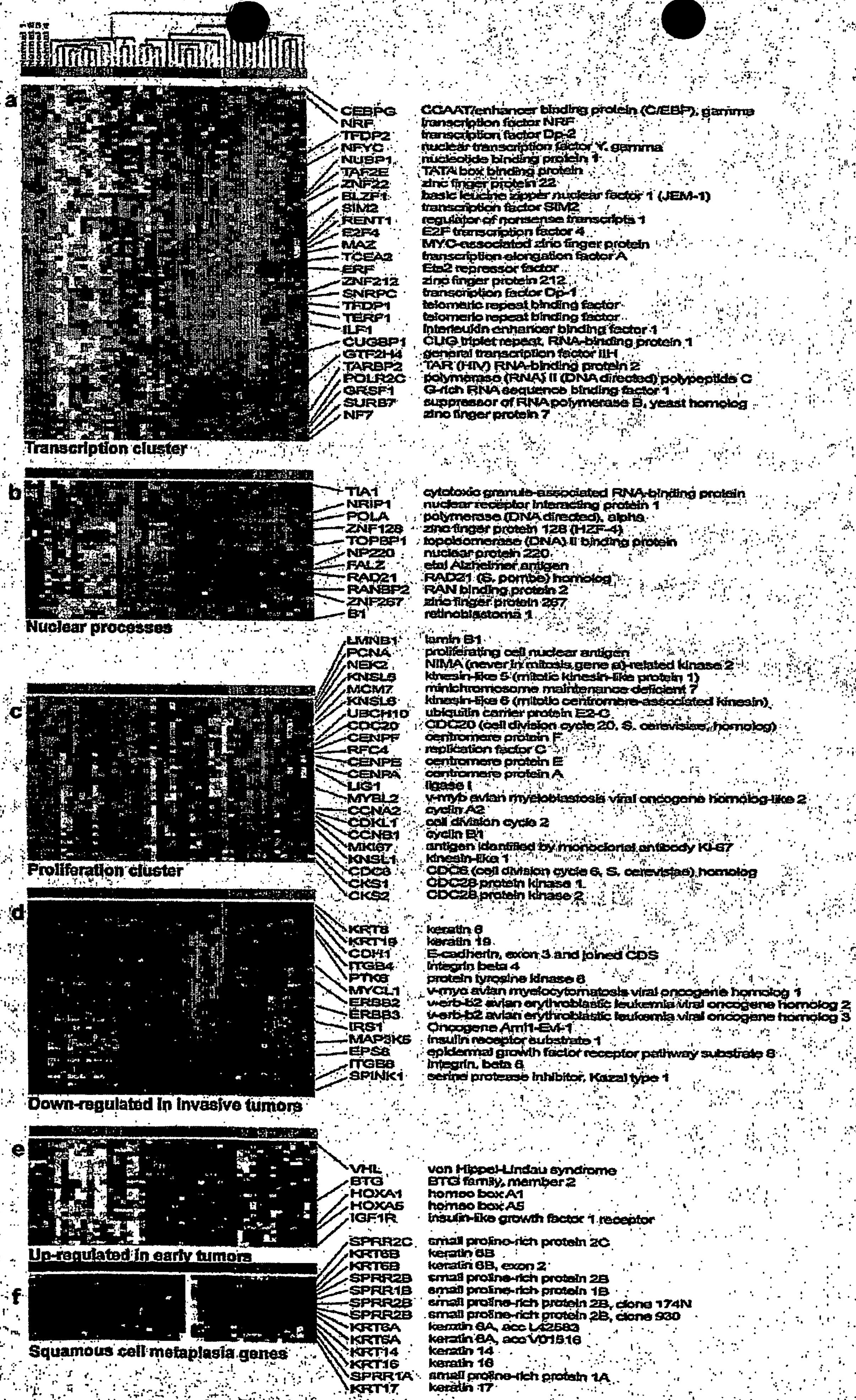
B	1327-1	Ta gr2	Papillary	no	5 month.
B	1352-1	Ta gr2	Papillary	no	6 month.
B	1379-1	Ta gr2	Papillary	no	5 month.
B	533-1	Ta gr2	Papillary	no	4 month.
B	679-1	Ta gr2	Papillary	no	4 month.
B	692-1	Ta gr2	Papillary	no	5 month.

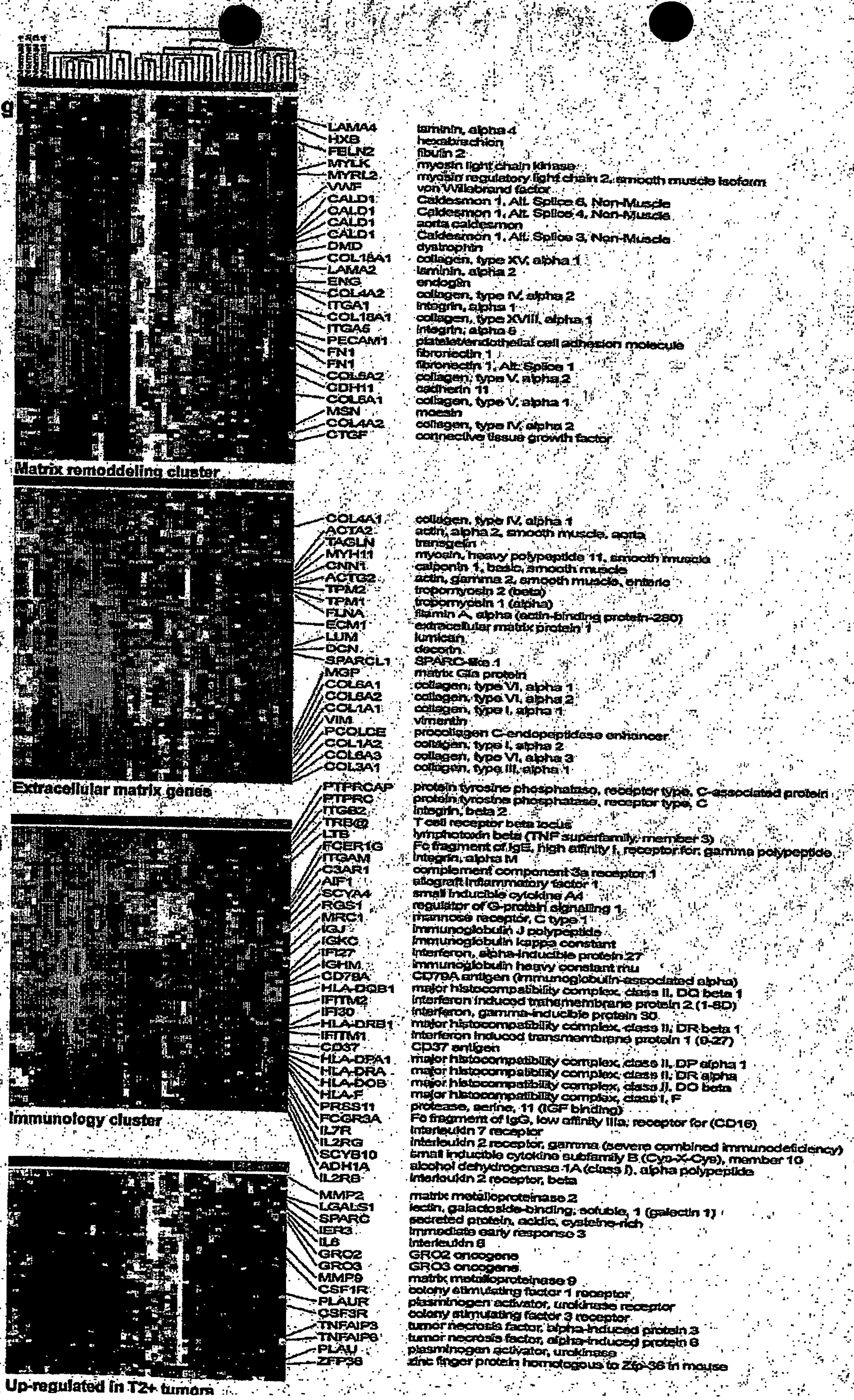
Group A: Primary tumours from patients with no recurrence of the disease for 2 years.

Group B: Primary tumours from patients with recurrence of the disease within 8 months.

Hierarchical cluster analysis results

Here we show expanded views of clusters a-j as identified in the 1767 gene-cluster. The tumour cluster dendrogram and colour bars on top of the clusters represents the same tumour cluster as shown in the paper. The four samples to the left are normal biopsies (normal 1-3) and a pool of 37 normal biopsies (normal 4).


a
CEBPG CCAAT/enhancer binding protein (C/EBP), gamma
NRF transcription factor NRF
TFDP2 transcription factor Dp-2
NFYC nuclear transcription factor Y, gamma
NUSP1 nucleotide binding protein 1
TAF2E TATA box binding protein
ZNF22 zinc finger protein 22
BLZF1 basic leucine zipper nuclear factor 1 (JEM-1)
SIM2 transcription factor SIM2
RENT1 regulator of nonsense transcripts 1
E2F4 E2F transcription factor 4
MAZ MYC-associated zinc finger protein
TCEA2 transcription elongation factor A
ERF Ets2 repressor factor
ZNF212 zinc finger protein 212
SNRPC transcription factor Dp-1
TFDP1 telomeric repeat binding factor
TERF1 telomeric repeat binding factor
ILF1 interleukin enhancer binding factor 1
CUGBP1 CUG triplet repeat, RNA-binding protein 1
GTF2H4 general transcription factor IIH
TARBP2 TAR (HIV) RNA-binding protein 2
POLR2C polymerase (RNA) II (DNA directed) polypeptide C
GRSF1 G-rich RNA sequence binding factor 1
SURB7 suppressor of RNA polymerase B, yeast homolog
NF7 zinc finger protein 7
Transcription cluster
b
TIA1 cytotoxic granule-associated RNA-binding protein
NRIP1 nuclear receptor interacting protein 1
POLA polymerase (DNA directed), alpha
ZNF128 zinc finger protein 128 (HZF-4)
TOPBP1 topoisomerase (DNA) II binding protein
NP220 nuclear protein 220
FALZ etal Alzheimer antigen
RAD21 RAD21 (S. pombe) homolog
RANBP2 RAN binding protein 2
ZNF267 zinc finger protein 267
B1 retinoblastoma 1
Nuclear processes
c
LMNB1 lamin B1
PCNA proliferating cell nuclear antigen
NEK2 NIMA (never in mitosis gene a)-related kinase 2
KNSL5 kinesin-like 5 (mitotic kinesin-like protein 1)
MCM7 minichromosome maintenance deficient 7
KNSL6 kinesin-like 6 (mitotic centromere-associated kinesin)
UBCH10 ubiquitin carrier protein E2-C
CDC20 CDC20 (cell division cycle 20, S. cerevisiae, homolog)
CENPF centromere protein F
RFC4 replication factor C
CENPE centromere protein E
CENPA centromere protein A
LIG1 ligase I
MYBL2 v-myb avian myeloblastosis viral oncogene homolog-like 2
CCNA2 cyclin A2
CDKL1 cell division cycle 2
CCNB1 cyclin B1
MKI67 antigen identified by monoclonal antibody Ki-67
KNSL1 kinesin-like 1
CDC6 CDC6 (cell division cycle 6, S. cerevisiae) homolog
CKS1 CDC28 protein kinase 1
CKS2 CDC28 protein kinase 2
Proliferation cluster
d
KRT8 keratin 8
KRT19 keratin 19
CDH1 E-cadherin, exon 3 and joined CDS
ITGB4 integrin beta 4
PTK6 protein tyrosine kinase 6
MYCL1 v-myc avian myelocytomatosis viral oncogene homolog 1
ERBB2 v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
ERBB3 v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3
IRS1 Oncogene Aml1-Evi-1
MAP3K5 insulin receptor substrate 1
EPS8 epidermal growth factor receptor pathway substrate 8
ITGB8 integrin, beta 8
SPINK1 serine protease inhibitor, Kazal type 1
Down-regulated in invasive tumors
e
VHL von Hippel-Lindau syndrome
BTG BTG family, member 2
HOXA1 homeo box A1
HOXA5 homeo box A5
IGF1R insulin-like growth factor 1 receptor
Up-regulated in early tumors
f
SPRR2C small proline-rich protein 2C
KRT6B keratin 6B
KRT6B keratin 6B, exon 2
SPRR2B small proline-rich protein 2B
SPRR1B small proline-rich protein 1B
SPRR2B small proline-rich protein 2B, clone 174N
SPRR2B small proline-rich protein 2B, clone 930
KRT6A keratin 6A, acc L42583
KRT6A keratin 6A, acc V01516
KRT14 keratin 14
KRT16 keratin 16
SPRR1A small proline-rich protein 1A
KRT17 keratin 17
Squamous cell metaplasia genes


g
LAMA4 laminin, alpha 4
HXB hexabrachion
FBLN2 fibulin 2
MYLK myosin light chain kinase
MYRL2 myosin regulatory light chain 2, smooth muscle isoform
VWF von Willebrand factor
CALD1 Caldesmon 1, Alt. Splice 6, Non-Muscle
CALD1 Caldesmon 1, Alt. Splice 4, Non-Muscle
CALD1 aorta caldesmon
CALD1 Caldesmon 1, Alt. Splice 3, Non-Muscle
DMD dystrophin
COL15A1 collagen, type XV, alpha 1
LAMA2 laminin, alpha 2
ENG endoglin
COL4A2 collagen, type IV, alpha 2
ITGA1 integrin, alpha 1
COL18A1 collagen, type XVIII, alpha 1
ITGA5 integrin, alpha 5
PECAM1 platelet/endothelial cell adhesion molecule
FN1 fibronectin 1
FN1 fibronectin 1, Alt. Splice 1
COL5A2 collagen, type V, alpha 2
CDH11 cadherin 11
COL5A1 collagen, type V, alpha 1
MSN moesin
COL4A2 collagen, type IV, alpha 2
CTGF connective tissue growth factor
Matrix remoddeling cluster
h
COL4A1 collagen, type IV, alpha 1
ACTA2 actin, alpha 2, smooth muscle, aorta
TAGLN transgelin
MYH11 myosin, heavy polypeptide 11, smooth muscle
CNN1 calponin 1, basic, smooth muscle
ACTG2 actin, gamma 2, smooth muscle, enteric
TPM2 tropomyosin 2 (beta)
TPM1 tropomyosin 1 (alpha)
FLNA filamin A, alpha (actin-binding protein-280)
ECM1 extracellular matrix protein 1
LUM lumican
DCN decorin
SPARCL1 SPARC-like 1
MGP matrix Gla protein
COL6A1 collagen, type VI, alpha 1
COL6A2 collagen, type VI, alpha 2
COL1A1 collagen, type I, alpha 1
VIM vimentin
PCOLCE procollagen C-endopeptidase enhancer
COL1A2 collagen, type I, alpha 2
COL6A3 collagen, type VI, alpha 3
COL3A1 collagen, type III, alpha 1
Extracellular matrix genes
i
PTPRCAP protein tyrosine phosphatase, receptor type, C-associated protein
PTPRC protein tyrosine phosphatase, receptor type, C
ITGB2 integrin, beta 2
TRB@ T cell receptor beta locus
LTB lymphotoxin beta (TNF superfamily, member 3)
FCER1G Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide
ITGAM integrin, alpha M
C3AR1 complement component 3a receptor 1
AIF1 allograft inflammatory factor 1
SCYA4 small inducible cytokine A4
RGS1 regulator of G-protein signalling 1
MRC1 mannose receptor, C type 1
IGJ immunoglobulin J polypeptide
IGKC immunoglobulin kappa constant
IFI27 interferon, alpha-inducible protein 27
IGHM immunoglobulin heavy constant mu
CD79A CD79A antigen (immunoglobulin-associated alpha)
HLA-DQB1 major histocompatibility complex, class II, DQ beta 1
IFITM2 interferon induced transmembrane protein 2 (1-8D)
IFI30 interferon, gamma-inducible protein 30
HLA-DRB1 major histocompatibility complex, class II, DR beta 1
IFITM1 interferon induced transmembrane protein 1 (9-27)
CD37 CD37 antigen
HLA-DPA1 major histocompatibility complex, class II, DP alpha 1
HLA-DRA major histocompatibility complex, class II, DR alpha
HLA-DOB major histocompatibility complex, class II, DO beta
HLA-F major histocompatibility complex, class I, F
PRSS11 protease, serine, 11 (IGF binding)
FCGR3A Fc fragment of IgG, low affinity IIIa, receptor for (CD16)
IL7R interleukin 7 receptor
IL2RG interleukin 2 receptor, gamma (severe combined immunodeficiency)
SCYB10 small inducible cytokine subfamily B (Cys-X-Cys), member 10
ADH1A alcohol dehydrogenase 1A (class I), alpha polypeptide
IL2RB interleukin 2 receptor, beta
Immunology cluster
j
MMP2 matrix metalloproteinase 2
LGALS1 lectin, galactoside-binding, soluble, 1 (galectin 1)
SPARC secreted protein, acidic, cysteine-rich
IER3 immediate early response 3
IL6 interleukin 6
GRO2 GRO2 oncogene
GRO3 GRO3 oncogene
MMP9 matrix metalloproteinase 9
CSF1R colony stimulating factor 1 receptor
PLAUR plasminogen activator, urokinase receptor
CSF3R colony stimulating factor 3 receptor
TNFAIP3 tumor necrosis factor, alpha-induced protein 3
TNFAIP6 tumor necrosis factor, alpha-induced protein 6
PLAU plasminogen activator, urokinase
ZFP36 zinc finger protein homologous to Zfp-36 in mouse
Up-regulated in T2+ tumors

Classification of samples

From the hierarchical cluster analysis of the samples (class discovery) we identified three major "molecular classes" of bladder carcinoma highly associated with the pathologic staging of the samples. Based on this finding we decided to build a molecular classifier that assigns tumours to these three "molecular classes". To build the classifier, we only used the tumours in which there was a correlation between the "molecular class" and the associated pathologic stage. Consequently, a T1 tumour clustering in the "molecular class" of T2 tumours was not used to build the classifier.

The genes used in the classifier were those genes with the highest values of the ratio (B/W) of the variation between the groups to the variation within the groups. High values of the ratio (B/W) signify genes with good group separation performance. We calculated the sum over the genes of the squared distance from the sample value to the group mean and classified the sample as belonging to the group where the distance to the group mean was smallest. If the relative difference between the distance to the closest and the second closest group compared to the distance to the closest group were below 5%, the classification failed and the sample was classified as belonging to both groups. The relative difference is refered to as the classifier strength.

Classifier performance

The classifier performance was tested using from 1-160 genes in cross-validation loops. Figure 1 shows that the closest correlation to histopathology is obtained in the cross-validation model using from 69-97 genes. Based on this we chose the model using 80 genes for cross-validation as our final classifier model.

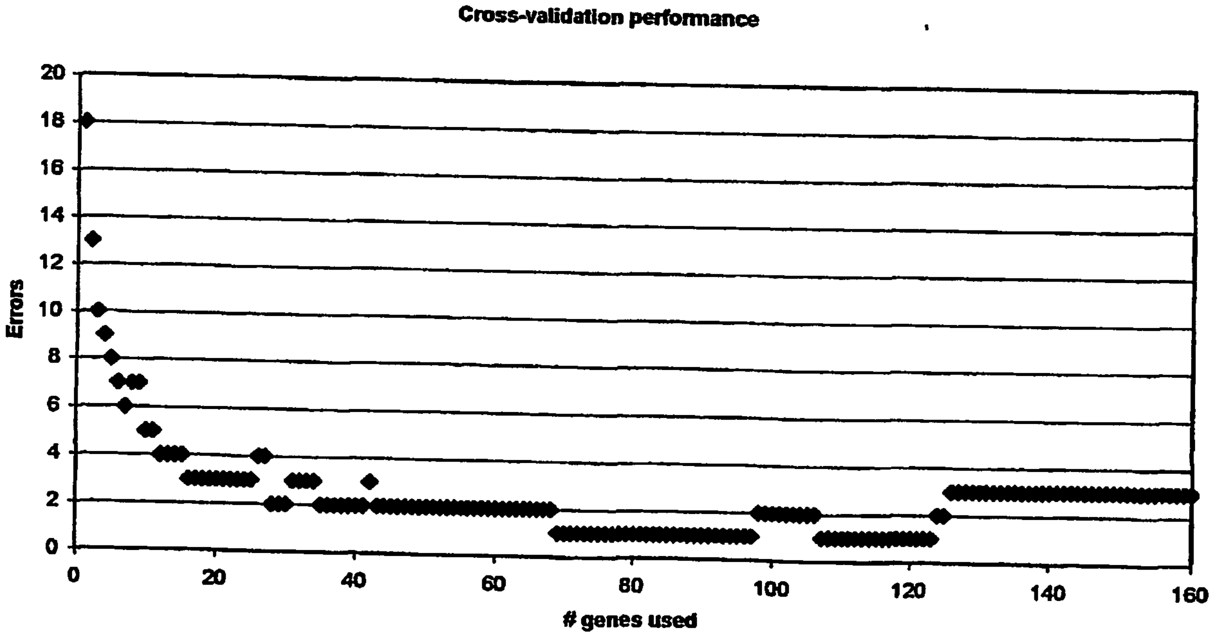


Figure 1 Number of classification errors vs. number of genes used in cross-validation loops.

Classifier model using 71 genes

We selected those genes for our final classifier model that were used in at least 75% (25 times) of the cross-validation loops. These 71 genes are listed in table 3.

Table 3 Feature: Accession number on HuGene fl array. Number: Number of times used in the 80 genes cross validation loops. Test (B/W): see below.

Feature	Unigene	Description	Number	Test (B/W)
AF000231_at	Hs.75618	RAB11A, member RAS oncogene family	33	26.77
D13666_s_at	Hs.136348	osteoblast specific factor 2 (fasciclin I-like)	33	27.71
D49372_s_at	Hs.54460	small inducible cytokine subfamily A (Cys-Cys), member 11	31	25.78
D83920_at	Hs.252136	ficolin (collagen/fibrinogen domain-containing) 1	33	31.18
D86479_at	Hs.118397	AE-binding protein 1	33	28.29
D89077_at	Hs.75367	Src-like-adaptor	33	30.03
D89377_at	Hs.89404	msh (Drosophila) homeo box homolog 2	33	51.50
HG4069-HT4339_s_at	-	Monocyte Chemotactic Protein 1	27	25.06
HG67-HT67_f_at	-	Zinc Finger Protein	33	27.81
HG907-HT907_at	-	Mg44	33	25.76
J02871_s_at	Hs.687	cytochrome P450, subfamily IVB, polypeptide 1	33	32.61
J03278_at	Hs.76144	platelet-derived growth factor receptor, beta polypeptide	33	28.02
J04058_at	Hs.169919	electron-transfer-flavoprotein, alpha polypeptide	33	29.46
J05032_at	Hs.80758	aspartyl-tRNA synthetase	33	38.21
J05070_at	Hs.151738	matrix metalloproteinase 9	33	35.34
J05448_at	Hs.79402	polymerase (RNA) II (DNA directed) polypeptide C (33kD)	32	26.51
K01396_at	Hs.297681	serine (or cysteine) proteinase inhibitor	33	28.66
L13720_at	Hs.78501	growth arrest-specific 6	33	29.69
M12125_at	Hs.300772	tropomyosin 2 (beta)	28	24.89
M15395_at	Hs.83968	integrin, beta 2	33	29.40
M16591_s_at	Hs.89555	hemopoietic cell kinase	33	32.34
M20530_at	-	pancreatic secretory trypsin inhibitor	33	30.28
M23178_s_at	Hs.73817	small inducible cytokine A3 (homologous to mouse Mip-1a)	33	35.36
M32011_at	Hs.949	neutrophil cytosolic factor 2	33	41.88
M33195_at	Hs.743	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide	33	30.40
M55998_s_at	-	alpha-1 collagen type I	33	26.83
M57731_s_at	Hs.75765	GRO2 oncogene	33	31.84
M68840_at	Hs.183109	monoamine oxidase A	33	32.39
M69203_s_at	-	small inducible cytokine A4	33	36.21
M72885_rna1_s_at	-	GOS2	33	27.94
M83822_at	Hs.62354	vesicle trafficking, beach and anchor containing	33	26.44
S77393_at	-	transcript ch138	33	49.85
U01833_at	Hs.81469	nucleotide binding protein 1 (E.coli MinD like)	33	30.62
U07231_at	Hs.309763	G-rich RNA sequence binding factor 1	33	39.10
U09937_rna1_s_at	-	urokinase-type plasminogen receptor	33	30.88
U10550_at	Hs.79022	GTP-binding protein overexpressed in skeletal muscle	28	25.26
U20158_at	Hs.2488	lymphocyte cytosolic protein 2	33	32.41
U41315_rna1_s_at	-	makorin, ring finger protein, 4	33	43.56
U47414_at	Hs.79069	cyclin G2	33	44.42
U49352_at	Hs.81548	2,4-dienoyl CoA reductase 1, mitochondrial	33	37.04
U50708_at	Hs.1265	branched chain keto acid dehydrogenase E1, beta polypeptide	33	42.89
U52101_at	Hs.9999	epithelial membrane protein 3	33	29.86
U64520_at	Hs.66708	vesicle-associated membrane protein 3 (cellubrevin)	33	30.17
U65093_at	Hs.82071	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	33	32.07
U68019_at	Hs.211578	MAD homolog 3	31	26.70
U68385_at	Hs.349772	Meis (mouse) homolog 3	33	31.56
U74324_at	Hs.90875	RAB interacting factor	33	30.26
U77970_at	Hs.321164	neuronal PAS domain protein 2	33	50.37
U90549_at	Hs.236774	high-mobility group protein 17-like 3	33	32.16
X04085_rna1_at	-	catalase	28	25.13
X07743_at	Hs.77436	pleckstrin	33	28.13
X13334_at	Hs.75627	CD14 antigen	33	35.79
X14046_at	Hs.153053	CD37 antigen	30	24.70
X15880_at	Hs.108885	collagen, type VI, alpha 1	33	31.51
X15882_at	-	collagen VI alpha-2 C-terminal globular domain	33	32.32
X51408_at	Hs.169965	chimerin 1	33	30.51

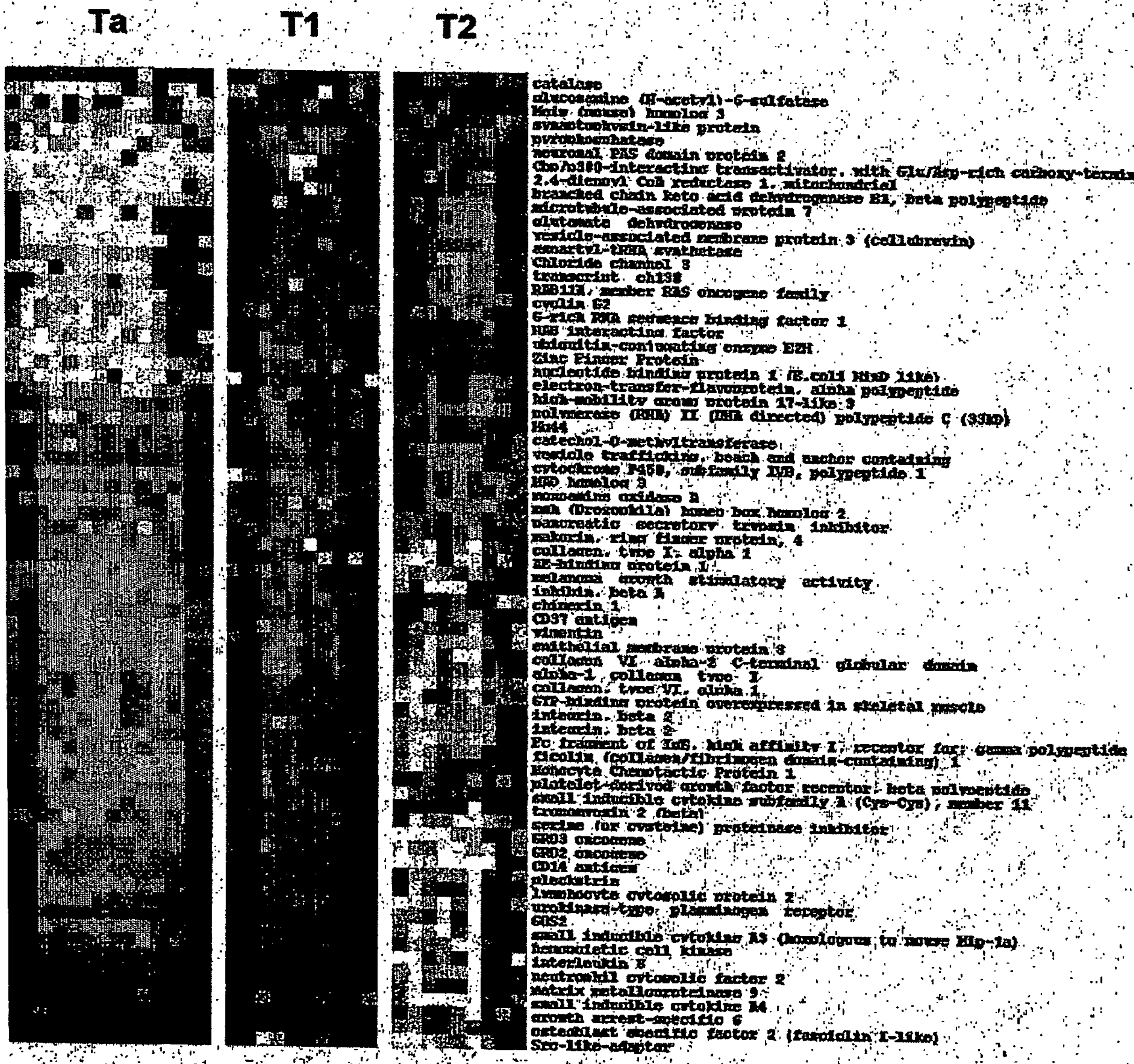
X53800_s_at	Hs.89690	GRO3 oncogene	33	33.63
X54489_ma1_at	-	melanoma growth stimulatory activity	33	33.57
X57579_s_at	Hs.727	inhibin, beta A	33	41.43
X64072_s_at	Hs.83968	integrin, beta 2	33	43.21
X67491_f_at	-	glutamate dehydrogenase	33	30.97
X68194_at	Hs.80919	synaptophysin-like protein	33	46.53
X73882_at	Hs.146388	microtubule-associated protein 7	33	53.16
X78520_at	Hs.174139	Chloride channel 3	33	47.38
Y00787_s_at	Hs.624	interleukin 8	32	27.54
Z12173_at	Hs.164036	glucosamine (N-acetyl)-6-sulfatase	30	25.44
Z19554_s_at	Hs.297753	vimentin	27	24.59
Z26491_s_at	Hs.240013	catechol-O-methyltransferase	32	26.92
Z29331_at	Hs.28505	ubiquitin-conjugating enzyme E2H	33	33.49
Z48605_at	-	pyrophosphatase	33	44.45
Z74615_at	Hs.172928	collagen, type I, alpha 1	33	55.18

Test for significance

To test the class separation performance of the 71 selected genes we compared the B/W ratios with the similar ratios of all the genes calculated from permutations of the arrays. For each permutation we construct three pseudogroups, pseudo-Ta, pseudo-T1, and pseudo-T2, so that the proportion of samples from the three original groups is approximately the same in the three pseudogroups. We then calculate the ratio of the variation between the psudogroups to the variation within the pseudogroups for all the genes. For 500 permutations we only two times had one gene for which the B/W value was higher than the lowest value for the original B/W values of the 71 selected genes (the two values being 25.28 and 25.93).

xpression profiles of the 71 genes classifie

The expression profiles of the 71 genes selected for our final classifier are shown in figure 2. The genes are clustered to obtain a better overview of similar expression patterns. From his it is obvious that the T1 stage is characterised by having expression patterns in common with either Ta or T2 tumours. There are no single genes that can be used as a T1 narker.



igure 2 Expression profiles of the 71 genes used in the final classifier model. The umours shown are the 33 tumours used in the cross-validation scheme. The Ta tumours re shown to the left, the T1 tumours in the middle, and the T2 tumours to the right.

20 genes for classifier (does not include previously published genes from our roup; *Cancer Res.2001.Mar.15.;61.(6.):2492.-9.* 61, 2492-2499))

hip accession number	**Accession number**
B000220_at	AB000220
C002073_cds1_at	AC002073

AF000231_at	AF000231
D10922_s_at	D10922
D10925_at	D10925
D11086_at	D11086
D11151_at	D11151
D13435_at	D13435
D13666_s_at	D13666
D14520_at	D14520
D21878_at	D21878
D26443_at	D26443
D28589_at	D28589
D42046_at	D42046
D45370_at	D45370
D49372_s_at	D49372
D50495_at	D50495
D63135_at	D63135
D64053_at	D64053
D83920_at	D83920
D85131_s_at	D85131
D86062_s_at	D86062
D86479_at	D86479
D86957_at	D86957
D86959_at	D86959
D86976_at	D86976
D87433_at	D87433
D87443_at	D87443
D87682_at	D87682
D89077_at	D89077
D89377_at	D89377
D90279_s_at	D90279
HG1996-HT2044_at	HG1996-HT2044
HG2090-HT2152_s_at	HG2090-HT2152
HG2463-HT2559_at	HG2463-HT2559
HG2994-HT4850_s_at	HG2994-HT4850
HG3044-HT3742_s_at	HG3044-HT3742
HG3187-HT3366_s_at	HG3187-HT3366
HG3342-HT3519_s_at	HG3342-HT3519
HG371-HT26388_s_at	HG371-HT26388
HG4069-HT4339_s_at	HG4069-HT4339
HG67-HT67_f_at	HG67-HT67
HG907-HT907_at	HG907-HT907
J02871_s_at	J02871
J03040_at	J03040
J03060_at	J03060
J03068_at	J03068
J03241_s_at	J03241
J03278_at	J03278
J03909_at	J03909
J03925_at	J03925
J04056_at	J04056
J04058_at	J04058

J04093_s_at	J04093	
J04130_s_at	J04130	
J04152_rna1_s_at	J04152	
J04162_at	J04162	
J04456_at	J04456	
J05032_at	J05032	
J05036_s_at	J05036	
J05070_at	J05070	
J05448_at	J05448	
K01396_at	K01396	
K03430_at	K03430	
L06797_s_at	L06797	
L10343_at	L10343	
L11708_at	L11708	
L13391_at	L13391	
L13698_at	L13698	
L13720_at	L13720	
L13923_at	L13923	
L15409_at	L15409	
L17325_at	L17325	
L19872_at	L19872	
L20971_at	L20971	
L22548_at	L22548	
L27476_at	L27476	
L29008_at	L29008	
L33799_at	L33799	
L40388_at	L40388	
L40904_at	L40904	
L41559_at	L41559	
L41919_rna1_at	L41919	
L42450_at	L42450	
L42621_at	L42621	
L43821_at	L43821	
L76465_at	L76465	
M11433_at	M11433	
M11718_at	M11718	
M11749_at	M11749	
M12125_at	M12125	
M13903_at	M13903	
M14058_at	M14058	
M14218_at	M14218	
M15395_at	M15395	
M16591_s_at		M16591
M16937_at	M16937	
M17219_at	M17219	
M19309_s_at		M19309
M19720_rna1_at	M19720	
M20530_at	M20530	
M23178_s_at		M23178
M24283_at	M24283	
M24486_s_at		M24486

M24902_at	M24902	
M27394_s_at		M27394
M27436_s_at		M27436
M28130_rna1_s_at	M28130	
M28211_at	M28211	
M29550_at	M29550	
M29971_at	M29971	
M31165_at	M31165	
M32011_at	M32011	
M33195_at	M33195	
M33374_at	M33374	
M37033_at	M37033	
M37766_at	M37766	
M55067_at	M55067	
M55153_at	M55153	
M55998_s_at		M55998
M57731_s_at		M57731
M58525_s_at		M58525
M59465_at	M59465	
M60278_at	M60278	
M62505_at	M62505	
M62840_at	M62840	
M63256_at	M63256	
M63262_at	M63262	
M64347_at	M64347	
M64925_at	M64925	
M68840_at	M68840	
M69066_at	M69066	
M69203_s_at		M69203
M72885_rna1_s_at	M72885	
M74719_at	M74719	
M77349_at	M77349	
M81118_at	M81118	
M82882_at	M82882	
M83652_s_at		M83652
M83822_at	M83822	
M84424_at	M84424	
M92934_at	M92934	
M95178_at	M95178	
M95787_at	M95787	
M98528_at	M98528	
M98539_at	M98539	
S49592_s_at	S49592	
S59049_at	S59049	
S62539_at	S62539	
S69115_at	S69115	
S77393_at	S77393	
S78187_at	S78187	
S82597_rna1_s_at	S82597	
S83325_s_at	S83325	
U01691_s_at	U01691	

U01833_at	U01833
U05227_at	U05227
U05861_at	U05861
U06681_at	U06681
U07231_at	U07231
U08021_at	U08021
U09278_at	U09278
U09578_at	U09578
U09770_at	U09770
U09937_rna1_s_at	U09937
U10099_s_at	U10099
U10550_at	U10550
U12424_s_at	U12424
U12535_at	U12535
U12778_at	U12778
U16306_at	U16306
U19713_s_at	U19713
U20158_at	U20158
U20536_s_at	U20536
U24266_at	U24266
U28249_at	U28249
U28488_s_at	U28488
U29680_at	U29680
U29953_rna1_at	U29953
U30313_at	U30313
U33818_at	U33818
U35735_at	U35735
U36341_rna1_at	U36341
U37143_at	U37143
U37431_at	U37431
U38175_at	U38175
U38864_at	U38864
U39840_at	U39840
U40490_at	U40490
U40705_at	U40705
U41060_at	U41060
U41315_rna1_s_at	U41315
U41745_at	U41745
U44111_at	U44111
U45878_s_at	U45878
U47414_at	U47414
U49352_at	U49352
U50534_at	U50534
U50708_at	U50708
U51010_s_at	U51010
U51711_at	U51711
U52101_at	U52101
U53003_at	U53003
U53225_at	U53225
U58046_s_at	U58046
U59913_at	U59913

U59914_at	U59914
U60205_at	U60205
U60975_at	U60975
U61981_at	U61981
U62389_at	U62389
U63289_at	U63289
U63824_at	U63824
U64520_at	U64520
U65093_at	U65093
U66619_at	U66619
U67156_at	U67156
U68019_at	U68019
U68385_at	U68385
U68485_at	U68485
U73514_at	U73514
U74324_at	U74324
U77970_at	U77970
U78027_rna4_at	U78027
U79271_at	U79271
U79751_at	U79751
U80456_at	U80456
U83303_cds2_at	U83303
U88871_at	U88871
U89942_at	U89942
U90549_at	U90549
U90716_at	U90716
U90916_at	U90916
U91985_at	U91985
V00594_at	V00594
V00594_s_at	V00594
X00371_rna1_at	X00371
X02761_s_at	X02761
X03663_at	X03663
X04011_at	X04011
X04085_rna1_at	X04085
X04500_at	X04500
X04602_s_at	X04602
X04741_at	X04741
X06256_at	X06256
X07203_at	X07203
X07438_s_at	X07438
X07743_at	X07743
X13334_at	X13334
X14046_at	X14046
X14813_at	X14813
X15306_rna1_at	X15306
X15573_at	X15573
X15880_at	X15880
X15882_at	X15882
X17042_at	X17042
X17644_s_at	X17644

X51408_at	X51408
X51757_at	X51757
X51823_at	X51823
X52022_at	X52022
X53331_at	X53331
X53800_s_at	X53800
X54489_rna1_at	X54489
X56687_s_at	X56687
X57351_s_at	X57351
X57579_s_at	X57579
X58072_at	X58072
X59770_at	X59770
X62048_at	X62048
X62466_at	X62466
X62535_at	X62535
X64044_at	X64044
X64072_s_at	X64072
X65614_at	X65614
X66945_at	X66945
X67491_f_at	X67491
X68194_at	X68194
X68314_at	X68314
X73882_at	X73882
X75042_at	X75042
X77794_at	X77794
X78520_at	X78520
X78549_at	X78549
X78565_at	X78565
X78669_at	X78669
X82209_at	X82209
X83572_at	X83572
X83618_at	X83618
X84908_at	X84908
X86098_at	X86098
X87241_at	X87241
X89109_s_at	X89109
X90908_at	X90908
X91504_at	X91504
X93036_at	X93036
X95097_rna1_s_at	X95097
X95592_at	X95592
X95632_s_at	X95632
X95677_at	X95677
X97267_rna1_s_at	X97267
Y00705_at	Y00705
Y00787_s_at	Y00787
Y00815_at	Y00815
Y07867_at	Y07867
Y08374_rna1_at	Y08374
Y12556_at	Y12556
Z12173_at	Z12173

Z19554_s_at	Z19554
Z22551_at	Z22551
Z26491_s_at	Z26491
Z29331_at	Z29331
Z35278_at	Z35278
Z35491_at	Z35491
Z48199_at	Z48199
Z48579_at	Z48579
Z48605_at	Z48605
Z74615_at	Z74615
Z74616_s_at	Z74616
Z79693_s_at	Z79693

Supervised learning prediction of recurrence

In this part of the work we identified genes differentially expressed between non-recurring and recurring tumours. Cross-validation and prediction was performed as previously described, except that genes are selected based on the value of the Wilcoxon statistic for difference between the two groups.

Prediction performance

The prediction perfomance was tested using from 1-200 genes in the cross-validation loops. Figure 3 below shows that the lowest error rate (8 errors) is obtained in e.g. the cross-validation model using from 39 genes. Based on this we selected this cross-validation model as our final predictor. The results of the predictions from the 39 gene cross-validation loops are listed in Table 6. The predictor misclassified four of the samples in each group and in one of the predictions the difference in the distances between the two group means is below the 5% difference limit as described above.

The probability of misclassifying 8 or less arrays by a random classification is 0.0053.

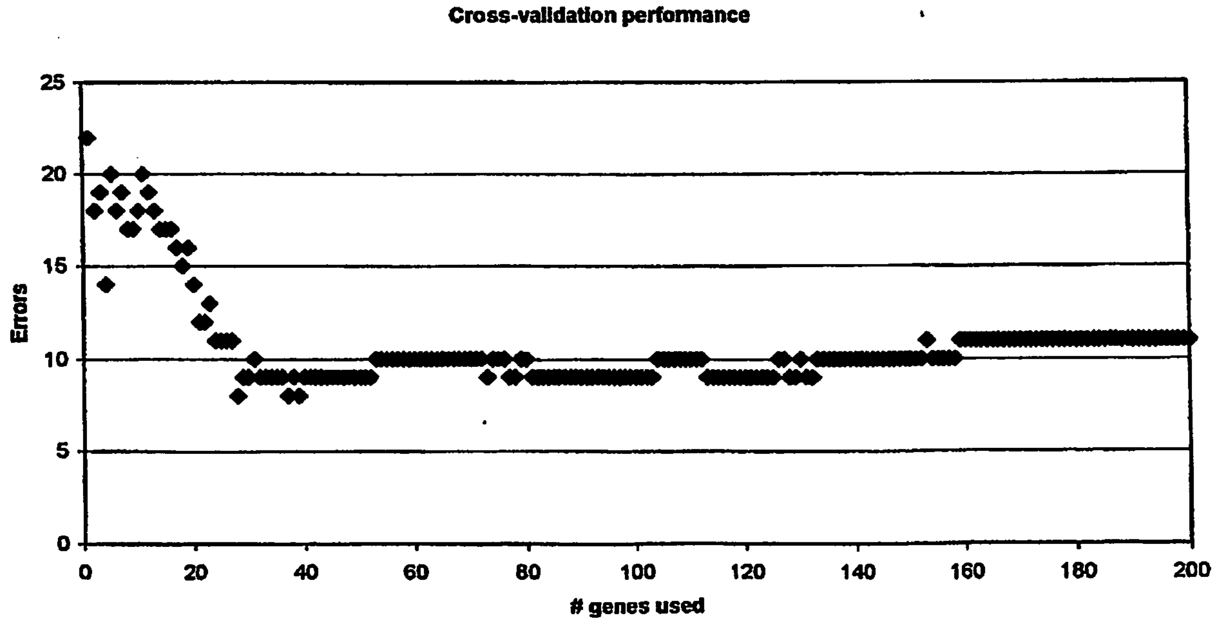


Figure 3 Number of prediction errors vs. number of genes used in cross-validation loops.

Table 6 Recurrence prediction results of 39 gene cross-validation loops. Group A: Primary tumours from patients with no recurrence of the disease for 2 years. Group B: Primary tumours from patients with recurrence of the disease within 8 months. Prediction, 0=no recurrence, 1=recurrence. Prediction strength: see p.8.

Group	Patient	Tumour (date)	Prediction	Error	Prediction strength
A	968-1	Ta gr2	0		0.19
A	928-1	Ta gr2	0		0.49
A	934-1	Ta gr2 (220798)	0		1.73
A	709-1	Ta gr2 (210798)	0		0.45
A	930-1	Ta gr2 (300698)	0		0.82
A	524-1	Ta gr2 (201095)	0		0.14
A	455-1	Ta gr2 (060695)	1	*	0.68
A	370-1	Ta gr2 (100195)	0		0.32
A	810-1	Ta gr2 (031097)	0		0.45
A	1146-1	Ta gr2 (231199)	0		0.98
A	1161-1	Ta gr2 (101299)	0		0.03
A	1006-1	Ta gr2 (231198)	1	*	1.57
A	942-1	Ta gr2	0		0.31
A	1060-1	Ta gr2	1	*	0.81
A	1255-1	Ta gr2	1	*	0.71
B	441-1	Ta gr2	1		1.03
B	780-1	Ta gr2	1		0.37
B	815-2	Ta gr2	1		0.35
B	829-1	Ta gr2	1		0.75
B	861-1	Ta gr2	0	*	2.55
B	925-1	Ta gr2	1		0.78
B	1008-1	Ta gr2	0	*	0.12
B	1086-1	Ta gr2	0	*	0.51
B	1105-1	Ta gr2	1		0.37
B	1145-1	Ta gr2	1		0.44
B	1327-1	Ta gr2	1		1.96
B	1352-1	Ta gr2	0	*	0.97
B	1379-1	Ta gr2	1		0.67
B	533-1	Ta gr2	1		0.31
B	679-1	Ta gr2	1		0.82
B	692-1	Ta gr2	1		0.45

Genes for classifier					
320 genes Chip accession numbers	160 genes	80 genes	40 genes	20 genes	10 ge
AB000220_at	AF000231_at	AF000231_at	D83920_at	D89377_at	D89377
AF000231_at	D13666_s_at	D13666_s_at	D89377_at	J05032_at	S7739
D10922_s_at	D21878_at	D49372_s_at	J02871_s_at	M23178_s_at	U41315_ma1_
D10925_at	D45370_at	D83920_at	J05032_at	M32011_at	U4741
D11086_at	D49372_s_at	D86479_at	J05070_at	M69203_s_at	U7797
D11151_at	D83920_at	D87433_at	M16591_s_at	S77393_at	X6819
D13435_at	D85131_s_at	D89077_at	M23178_s_at	U07231_at	X7388
D13666_s_at	D86062_s_at	D89377_at	M32011_at	U41315_ma1_s_at	X7852
D14520_at	D86479_at	HG4069-HT4339_s_at	M33195_at	U47414_at	Z4860
D21878_at	D86957_at	HG67-HT67_f_at	M57731_s_at	U49352_at	Z7461
D26129_at	D86976_at	HG907-HT907_at	M68840_at	U50708_at	
D26443_at	D87433_at	J02871_s_at	M69203_s_at	U77970_at	
D42046_at	D89077_at	J03278_at	S77393_at	X13334_at	
D42047_at	D89377_at	J04058_at	U01833_at	X57579_s_at	
D45370_at	HG3044-HT3742_s_at	J05032_at	U07231_at	X64072_s_at	
D49372_s_at	HG371-HT26388_s_at	J05070_at	U09937_ma1_s_at	X68194_at	
D49387_at	HG4069-HT4339_s_at	J05448_at	U20158_at	X73882_at	
D50495_at	HG67-HT67_f_at	K01396_at	U41315_ma1_s_at	X78520_at	
D63135_at	HG907-HT907_at	L13720_at	U47414_at	Z48605_at	
D64053_at	J02871_s_at	L40904_at	U49352_at	Z74615_at	
D83920_at	J03040_at	M12125_at	U50708_at		
D85131_s_at	J03068_at	M15395_at	U65093_at		
D86062_s_at	J03241_s_at	M16591_s_at	U68385_at		
D86479_at	J03278_at	M20530_at	U77970_at		
D86957_at	J03909_at	M23178_s_at	U90549_at		
D86959_at	J04058_at	M32011_at	X13334_at		
D86974_at	J04130_s_at	M33195_at	X15880_at		
D86976_at	J04162_at	M55998_s_at	X15882_at		
D87120_at	J04456_at	M57731_s_at	X51408_at		
D87433_at	J05032_at	M63262_at	X53800_s_at		
D87443_at	J05070_at	M68840_at	X54489_ma1_at		
D87682_at	J05448_at	M69203_s_at	X57579_s_at		
D89077_at	K01396_at	M72885_ma1_s_at	X64072_s_at		
D89377_at	K03430_at	M83822_at	X67491_f_at		
D90279_s_at	L13698_at	S77393_at	X68194_at		
HG1996-	L13720_at	U01833_at	X73882_at		

HT2044_at
HG2090-HT2152_s_at
HG2379-HT3996_s_at
HG2463-HT2559_at
HG2724-HT2820_at
HG3044-HT3742_s_at
HG3187-HT3366_s_at
HG3342-HT3519_s_at
HG371-HT26388_s_at
HG4069-HT4339_s_at
HG67-HT67_f_at
HG907-HT907_at
J02871_s_at
J03040_at
J03060_at
J03068_at
J03241_s_at
J03278_at
J03909_at
J03925_at
J04056_at
J04058_at
J04130_s_at
J04152_rna1_s_at
J04162_at
J04456_at
J05032_at
J05070_at
J05448_at
K01396_at
K03430_at
L06797_s_at
L07956_at
L10343_at
L11672_r_at
L13391_at
L13698_at
L13720_at
L13923_at
L15409_at

L13923_at
L15409_at
L17325_at
L19872_at
L27476_at
L33799_at
L40388_at
L40904_at
L41919_rna1_at
M11433_at
M11718_at
M12125_at
M14218_at
M15395_at
M16591_s_at
M17219_at
M20530_at
M23178_s_at
M28130_rna1_s_at
M29550_at
M31165_at
M32011_at
M33195_at
M37033_at
M37766_at
M55998_s_at
M57731_s_at
M62840_at
M63262_at
M68840_at
M69203_s_at
M72885_rna1_s_at
M77349_at
M82882_at
M83822_at
M92934_at
M95178_at
S69115_at
S77393_at

U07231_at
U09937_rna1_s_at
U10550_at
U20158_at
U28488_s_at
U29680_at
U41315_rna1_s_at
U47414_at
U49352_at
U50708_at
U52101_at
U59914_at
U64520_at
U65093_at
U68019_at
U68385_at
U74324_at
U77970_at
U90549_at
X04085_rna1_at
X07438_s_at
X07743_at
X13334_at
X14046_at
X15880_at
X15882_at
X51408_at
X53800_s_at
X54489_rna1_at
X57579_s_at
X62048_at
X64072_s_at
X67491_f_at
X68194_at
X73882_at
X78520_at
X97267_rna1_s_at
Y00787_s_at
Z12173_at

X78520_at
Z29331_at
Z48605_at
Z74615_at

L17325_at
L19872_at
L20971_at
L22548_at
L25444_at
L27476_at
L29008_at
L33799_at
L40388_at
L40904_at
L41559_at
L41919_rna1_at
L42450_at
L42621_at
L43821_at
M11433_at
M11718_at
M11749_at
M12125_at
M14058_at
M14218_at
M15395_at
M16591_s_at
M16937_at
M17219_at
M19309_s_at
M19720_rna1_at
M20530_at
M23178_s_at
M24283_at
M24902_at
M27394_s_at
M27436_s_at
M28130_rna1_s_at
M28211_at
M29550_at
M29971_at
M31165_at
M32011_at
M33195_at
M33374_at
M34309_at
M37033_at
M37766_at
M55067_at
M55153_at

S78187_at
U01833_at
U07231_at
U09278_at
U09937_rna1_s_at
U10550_at
U12424_s_at
U16306_at
U20158_at
U20536_s_at
U24266_at
U28249_at
U28488_s_at
U29680_at
U37143_at
U38864_at
U39840_at
U41315_rna1_s_at
U44111_at
U47414_at
U49352_at
U50708_at
U52101_at
U59914_at
U60205_at
U61981_at
U64520_at
U65093_at
U66619_at
U68019_at
U68385_at
U68485_at
U74324_at
U77970_at
U83303_cds2_at
U88871_at
U90549_at
U90716_at
V00594_at
V00594_s_at
X02761_s_at
X04011_at
X04085_rna1_at
X07438_s_at
X07743_at
X13334_at

Z19554_s_at
Z26491_s_at
Z29331_at
Z48605_at
Z74615_at

M55998_s_at
M57731_s_at
M59465_at
M60278_at
M62505_at
M62840_at
M63256_at
M63262_at
M64925_at
M68840_at
M69066_at
M69203_s_at
M72885_rna1_s_at
M74719_at
M77349_at
M81118_at
M82882_at
M83652_s_at
M83822_at
M92934_at
M93426_at
M95178_at
M95787_at
M98528_at
M98539_at
S49592_s_at
S59049_at
S62539_at
S69115_at
S77393_at
S78187_at
S83325_s_at
U01691_s_at
U01833_at
U03851_at
U05227_at
U05861_at
U06681_at
U07231_at
U08021_at
U09278_at
U09578_at
U09770_at
U09937_rna1_s_at
U10099_s_at
U10550_at

X14046_at
X14813_at
X15880_at
X15882_at
X51408_at
X53800_s_at
X54489_rna1_at
X57351_s_at
X57579_s_at
X58072_at
X62048_at
X64072_s_at
X65614_at
X66945_at
X67491_f_at
X68194_at
X73882_at
X78520_at
X78549_at
X78565_at
X78669_at
X83618_at
X84908_at
X90908_at
X91504_at
X95632_s_at
X97267_rna1_s_at
Y00705_at
Y00787_s_at
Y00815_at
Y08374_rna1_at
Z12173_at
Z19554_s_at
Z26491_s_at
Z29331_at
Z35491_at
Z48199_at
Z48605_at
Z74615_at

U12424_s_at
U12535_at
U12778_at
U16306_at
U19713_s_at
U20158_at
U20536_s_at
U24266_at
U24577_at
U28249_at
U28368_at
U28488_s_at
U29680_at
U29953_ma1_at
U30313_at
U33818_at
U35735_at
U36341_ma1_at
U37143_at
U37431_at
U38175_at
U38864_at
U39840_at
U40490_at
U40705_at
U41315_ma1_s_at
U41745_at
U42360_cds2_at
U44111_at
U45878_s_at
U46461_at
U47414_at
U49352_at
U50534_at
U50708_at
U51010_s_at
U51711_at
U52101_at
U52960_at
U53003_at
U53225_at
U58046_s_at
U59913_at
U59914_at
U60205_at
U61981_at
U62389_at
U63289_at
U63824_at
U64520_at

U65093_at
U66619_at
U68019_at
U68385_at
U68485_at
U70063_at
U73514_at
U74324_at
U77970_at
U78027_rna4_at
U79271_at
U79751_at
U80456_at
U83303_cds2_at
U88871_at
U89942_at
U90549_at
U90716_at
U91985_at
V00594_at
V00594_s_at
X00371_rna1_at
X02761_s_at
X03663_at
X04011_at
X04085_rna1_at
X04500_at
X04602_s_at
X04741_at
X06256_at
X07203_at
X07438_s_at
X07743_at
X12530_s_at
X13334_at
X14046_at
X14813_at
X15306_rna1_at
X15573_at
X15880_at
X15882_at
X17042_at
X17644_s_at
X51408_at
X51757_at
X51823_at
X52022_at
X53331_at
X53800_s_at
X54489_rna1_at
X56687_s_at

X57351_s_at
X57579_s_at
X58072_at
X59770_at
X62048_at
X62466_at
X62535_at
X64044_at
X64072_s_at
X65614_at
X66945_at
X67491_f_at
X68194_at
X73882_at
X75042_at
X78520_at
X78549_at
X78565_at
X78669_at
X82209_at
X83572_at
X83618_at
X84908_at
X86098_at
X89109_s_at
X90858_at
X90908_at
X91504_at
X93036_at
X95097_rna1_s_a
t
X95592_at
X95632_s_at
X95677_at
X97267_rna1_s_a
t
Y00705_at
Y00787_s_at
Y00815_at
Y07867_at
Y08374_rna1_at
Y12556_at
Z12173_at
Z19554_s_at
Z26491_s_at
Z29331_at
Z35278_at
Z35491_at
Z48199_at
Z48579_at
Z48605_at

Z74615_at
Z74616_s_at
Z79693_s_at

26 gene recurrence predictor

We selected the genes used in at least 29 of the 31 cross-validation loops to constitute our final recurrence prediction model. These 26 genes are listed in table 7.

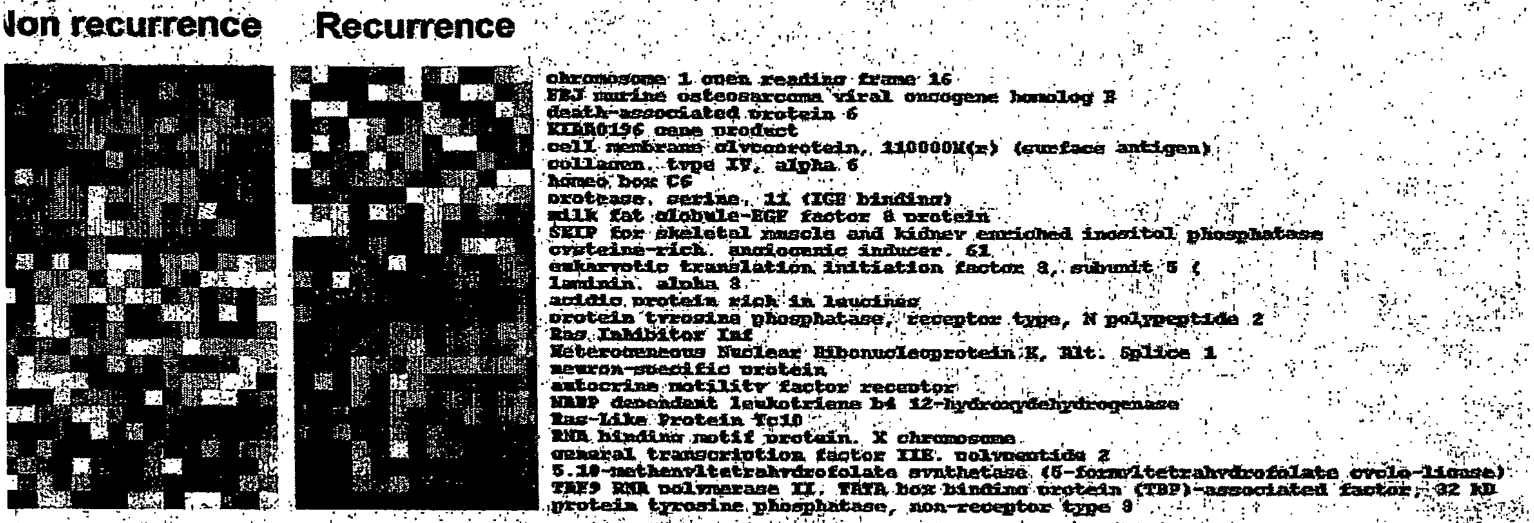
Table 7 The 26 genes that we find optimal for recurrence prediction.

Feature	Unigene	Description	Number	Test (W-N)
AF006041_at	Hs.336916	death-associated protein 6	31	0.054 (161-7)
D21337_at	Hs.408	collagen, type IV, alpha 6	31	0.058 (160-6)
D49387_at	-	NADP dependent leukotriene b4 12-hydroxydehydrogenase	31	0.118 (313-8)
D64154_at	Hs.90107	cell membrane glycoprotein, 110000M(r) (surface antigen).	31	0.078 (165-9)
D83780_at	Hs.8294	KIAA0196 gene product	31	0.094 (159-4)
D87258_at	Hs.75111	protease, serine, 11 (IGF binding)	30	0.112 (168-11)
D87437_at	Hs.15087	chromosome 1 open reading frame 16	31	0.058 (160-6)
HG1879-HT1919_at	-	Ras-Like Protein Tc10	31	0.122 (314-7)
HG3076-HT3238_s_at	-	Heterogeneous Nuclear Ribonucleoprotein K, Alt. Splice 1	31	0.080 (309-17)
HG511-HT511_at	-	Ras Inhibitor Inf	31	0.348 (319-2)
L34155_at	Hs.83450	laminin, alpha 3	31	0.122 (314-7)
L38928_at	Hs.118131	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)	29	0.348 (319-2)
L49169_at	Hs.75678	FBJ murine osteosarcoma viral oncogene homolog B	31	0.108 (155-2)
M16938_s_at	Hs.820	homeo box C6	29	0.09 (170-16)
M63175_at	Hs.80731	autocrine motility factor receptor	29	0.098 (308-18)
M64572_at	Hs.153932	protein tyrosine phosphatase, non-receptor type 3	31	0.064 (305-31)
M98528_at	Hs.79404	neuron-specific protein	31	0.122 (314-7)
U21858_at	Hs.60679	TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 32 kD	31	0.122 (314-7)
U45973_at	Hs.178347	SKIP for skeletal muscle and kidney enriched inositol phosphatase	31	0.094 (310-14)
U58516_at	Hs.3745	milk fat globule-EGF factor 8 protein	29	0.100 (175-28)
U62015_at	Hs.8867	cysteine-rich, angiogenic inducer, 61	31	0.106 (169-13)
U66702_at	Hs.74624	protein tyrosine phosphatase, receptor type, N polypeptide 2	31	0.146 (149-1)
U70439_s_at	Hs.84264	acidic protein rich in leucines	30	0.08 (309-17)
U94855_at	Hs.7811	eukaryotic translation initiation factor 3, subunit 5 (	30	0.092 (311-12)
X63469_at	Hs.77100	general transcription factor IIE, polypeptide 2	31	0.092 (311-12)
Z23064_at	Hs.146381	RNA binding motif protein, X chromosome	30	0.066 (307-24)

Number: Number of times the gene has been used in a cross-validation loop. Test: The numbers in parenthesis are the value W of the Wilcoxon test statistic for no difference between the two groups together with the number N of genes for which the Wilcoxon test statistic is bigger than or equal to the value W. The test value is obtained from 500 permutations of the arrays. In each permutation we form new pseudogroups where both of the pseudogroups have the same proportion of arrays from the two original groups. For each permutation we count the number of genes for which the Wilcoxon test statistic based on the pseudogroups is bigger than or equal to W, and the test value is the proportion of the permutations for which this number is bigger than or equal to N. Thus the test value measures the significance of the observed value W. Consequently, for most of our selected genes we only find as least as good predictive genes in about 10% of the formed pseudogroups.

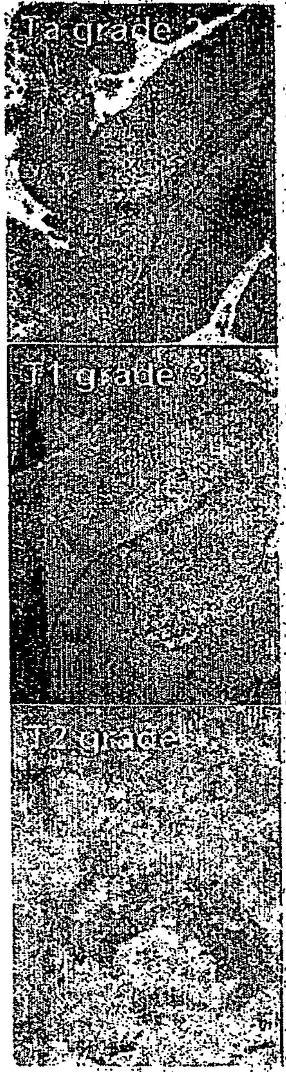
xpression profiles of the 26 [illegible]e recurrence predictor

he expression profiles of the 26 genes that were used in more than 75% of the cross-validation loops are hown in figure 4 below.



igure 4 The expression profiles of the 26 genes that constitute our final prediction model. The genes are listed according to the degree of correlation with the recurrence and non-ecurrence groups. Genes with highest correlations are found in the top and the bottom of he list.

Table 1 Clinical data on dise[illegible]ourses and results from molecular [illegible]sification

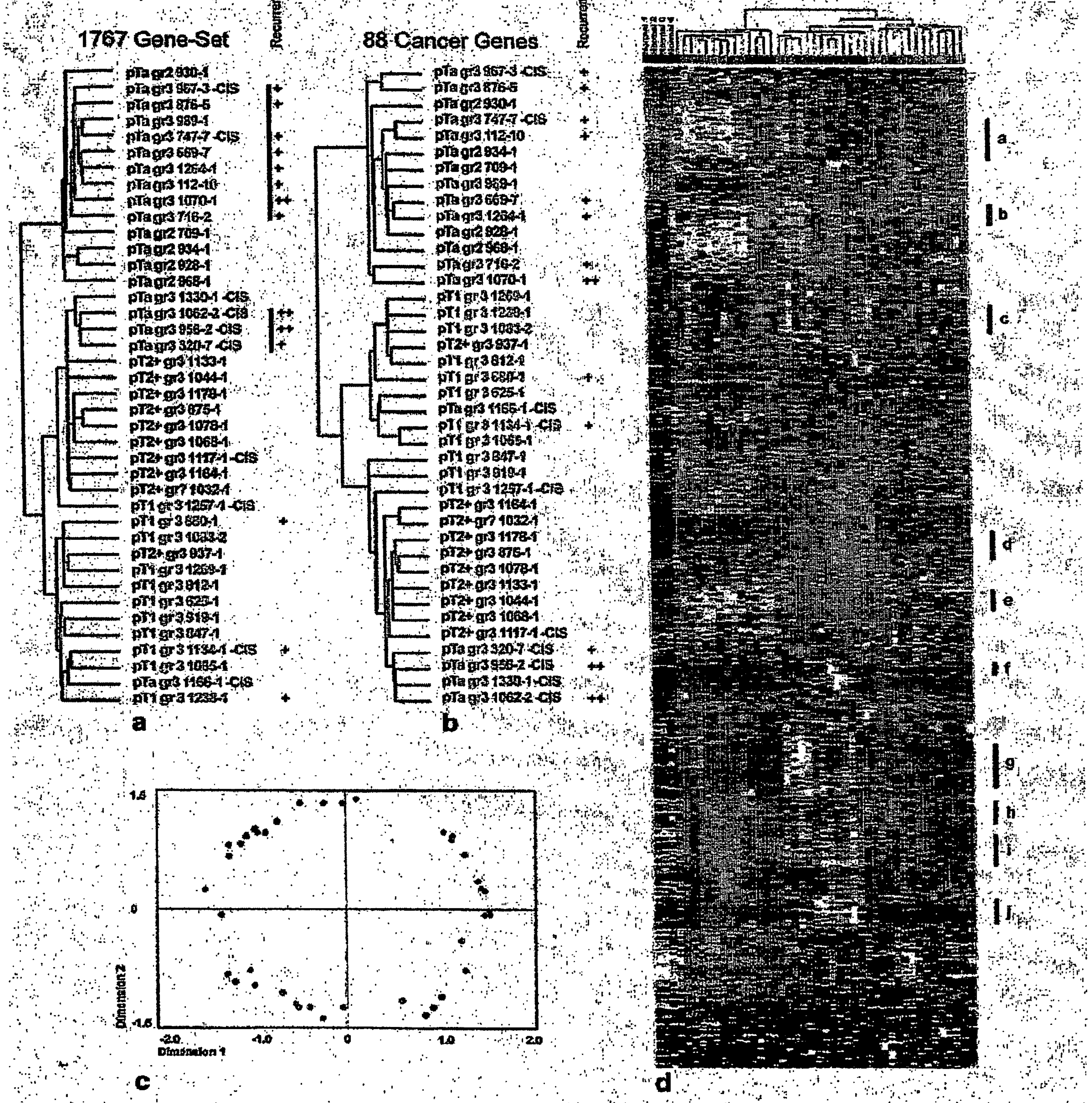
Tumours*	Patient	Previous tumours	Tumour analysed	Subsequent tumours	Carcinoma *in situ*†	Reviewed histology‡	Molecular classifier§		
							320	80	20
	Ta grade II tumours – no progression								
	709-1		Ta gr2		No	Ta gr3	Ta	Ta	Ta
	968-1		Ta gr2	1 Ta	No		Ta/T1	Ta	Ta
	934-1		Ta gr2		No		T1	Ta	Ta
	928-1		Ta gr2		No		Ta	Ta	T1
	930-1		Ta gr2		No		Ta	Ta	Ta
	Ta grade III tumours – no prior T1 tumour or CIS								
	989-1		Ta gr3		No		Ta	Ta	Ta
	1264-1		Ta gr3	3 Ta	No		Ta	Ta	Ta
	876-5	4 Ta	Ta gr3		No		Ta	Ta	Ta
	669-7	5 Ta	Ta gr3	4 Ta	No	Ta gr2	Ta	Ta	Ta
	716-2	1 Ta	Ta gr3	2 Ta	No		Ta	Ta	Ta
	Ta grade III tumours – no prior T1 tumour but CIS in selected site biopsies								
	1070-1		Ta gr3	1 Ta	Subsequent visit		Ta	Ta	Ta
	956-2		Ta gr3	1 Ta	Sampling visit		T2	T2	T2/T1
	1062-2		Ta gr3	1 T1	Sampling visit		T2/Ta	T1/Ta	Ta
	1166-1		Ta gr3		Sampling visit		Ta/T1	Ta	Ta
	1330-1		Ta gr3		Sampling visit		T2	T2	Ta
	Ta grade III tumours – a prior T1 tumour and CIS in selected site biopsies								
	747-7	5 Ta, 1 T1	Ta gr3	3 Ta	Sampling visit		Ta	Ta	Ta
	112-10	7 Ta, 2 T1	Ta gr3	2 Ta, 4 T1	Previous visit		Ta	Ta	Ta
	320-7	1 Ta, 2 T1	Ta gr3	2 Ta	Sampling visit		T2	T2	Ta
	967-3	2 T1	Ta gr3	1 T1	Sampling visit		Ta	Ta	Ta
	T1 grade III tumours – no prior muscle invasive tumour								
	625-1		T1 gr3		No		T1	T1	T1
	847-1		T1 gr3		No		T1	T1	T1
	1257-1		T1 gr3		Sampling visit		T1	T1	T1
	919-1		T1 gr3		No		T1	T1	T1
	880-1		T1 gr3	4 Ta	No		T1	T1	T1
	812-1		T1 gr3		No		T1	T1	T1
	1269-1		T1 gr3		No	No review	T1	T1	T1
	1083-2	1 Ta	T1 gr3		No	No review	T1	T1	T1
	1238-1		T1 gr3	1 Ta, 1 T2+	No		T1	T1	T1
	1065-1		T1 gr3		Subsequent visit	No review	T1	T1	T1
	1134-1		T1 gr3	3 T1	Sampling visit	T2 gr3	T1	T1	T1
	T2+ grade III/IV tumours – only primary tumours								
	1164-1		T2+ gr4		No	T2+ gr3	T2/T1	T1	T1
	1032-1		T2+ gr?		ND	No review	T2	T2	T2
	1117-1		T2+ gr3		ND		T2	T2	T1
	1178-1		T2+ gr3		ND		T2	T2	T2
	1078-1		T2+ gr3		ND		T2	T2	T2
	875-1		T2+ gr3		No		T2	T2	T2
	1044-1		T2+ gr3	1 T2+	ND		T2	T2	T2
	1133-1		T2+ gr3		ND		T2	T2	T2
	1068-1		T2+ gr3		No		T2	T2	T2
	937-1		T2+ gr3		ND	No review	T1	T1	T1

* Examples of tumour histology.

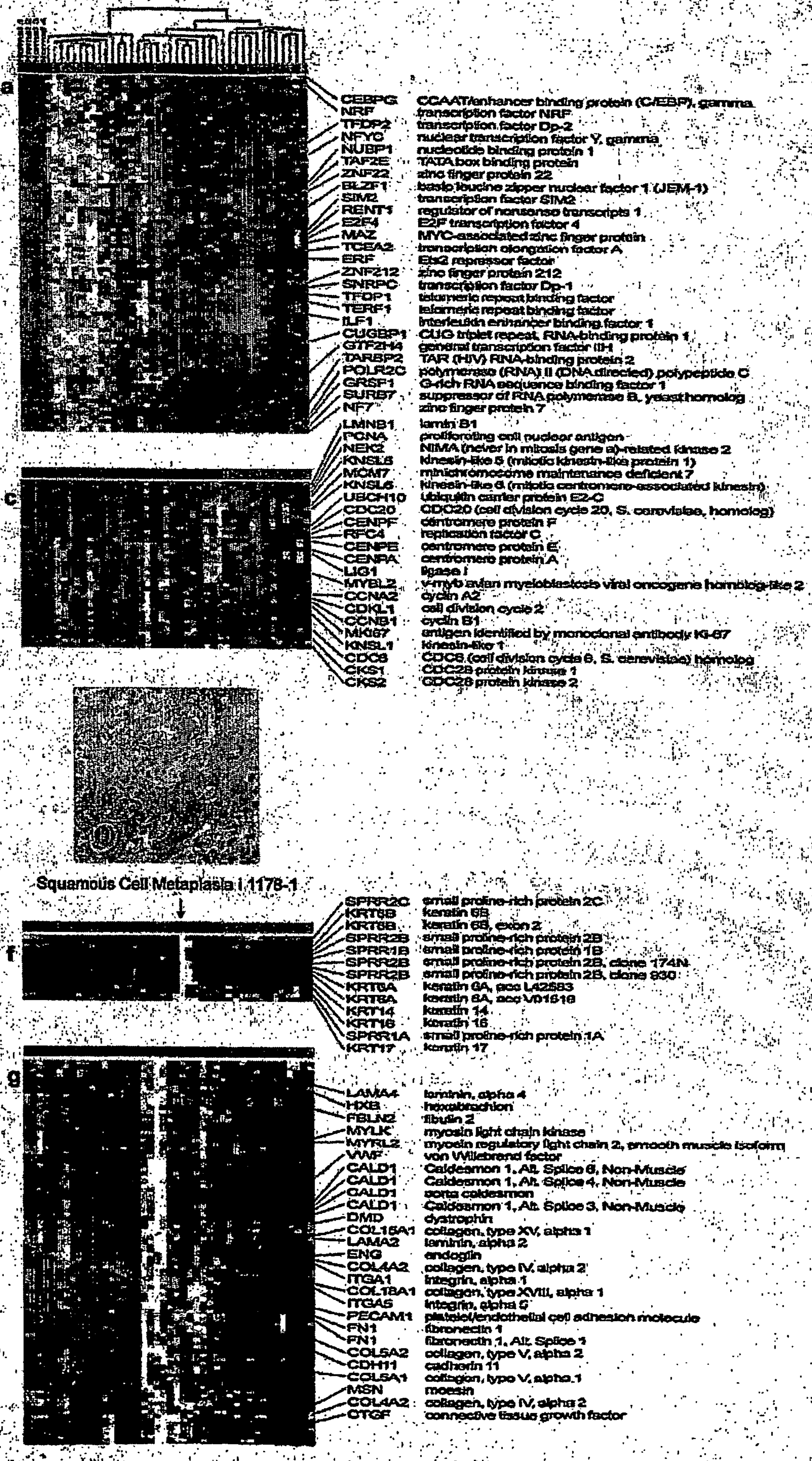
† Carcinoma *in situ* detected in selected site biopsies at the time of sampling tumour tissue for the arrays or at previous or subsequent visits.

‡ All tumours were reviewed by a single uro-pathologist and any change compared to the routine classification is listed.

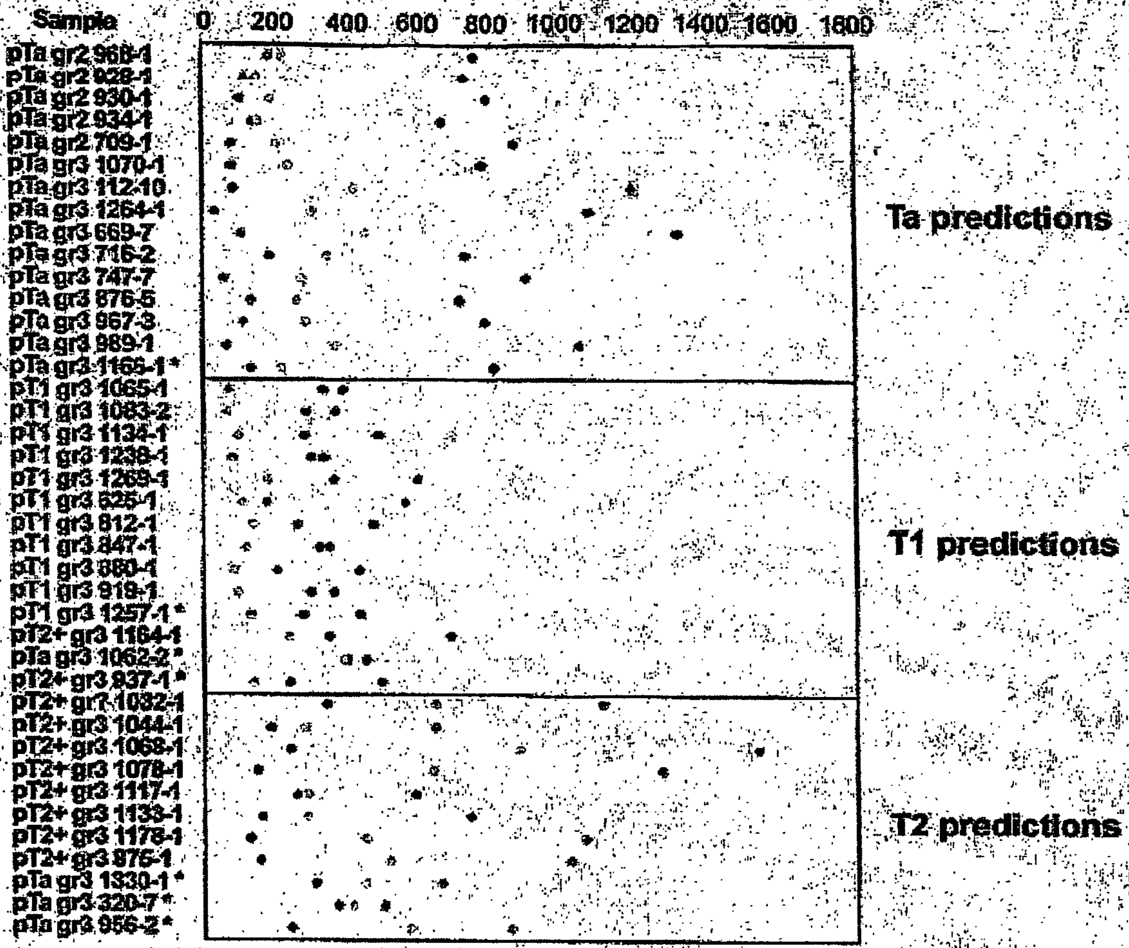
§ Molecular classification based on 320, 80, and 20 genes respectively.



Ørntoft_fig1



Ørntoft_fig2



Ørntoft_fig3

Expression profiles of the 26-gene recurrence predictor

The expression profiles of the 26 genes that were used in more than 75% of the cross-validation loops are shown in figure 4 below.

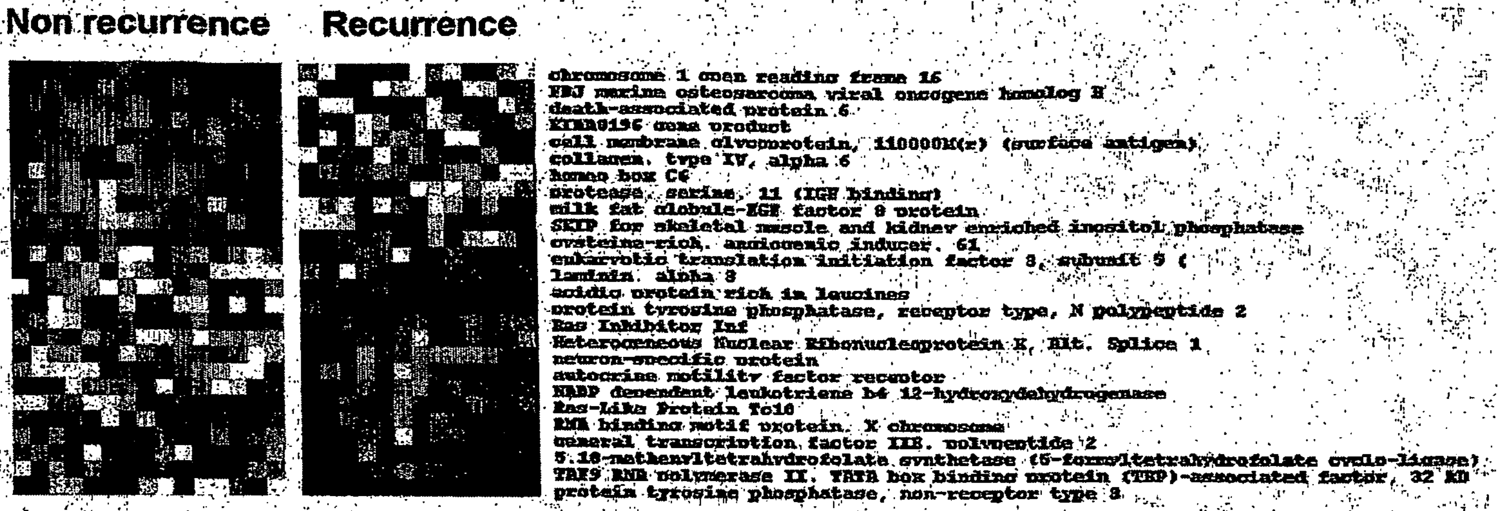


Figure 4 The expression profiles of the 26 genes that constitute our final prediction model. The genes are listed according to the degree of correlation with the recurrence and non-recurrence groups. Genes with highest correlations are found in the top and the bottom of the list.

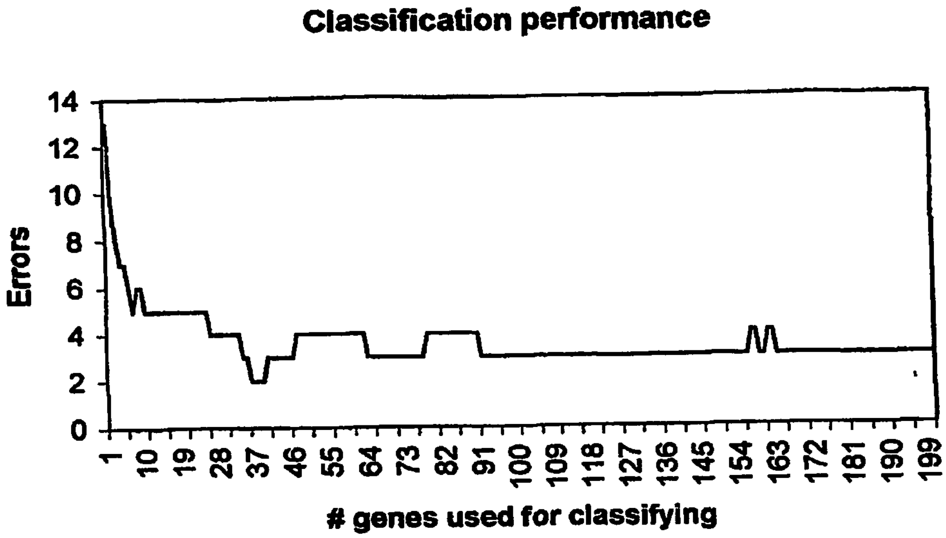
Web Table B: Patient disease course information – recurrence vs. no recurrence

Group	Patient	Tumour (date)	Pattern	Carcinoma *in situ*	Time to recurrence
A	968-1	Ta gr2	Papillary	no	27 month
A	928-1	Ta gr2	Papillary	no	38 month.
A	934-1	Ta gr2 (220798)	Papillary	no	-
A	709-1	Ta gr2 (210798)	Papillary	no	-
A	930-1	Ta gr2 (300698)	Papillary	no	-
A	524-1	Ta gr2 (201095)	Papillary	no	-
A	455-1	Ta gr2 (060695)	Papillary	no	-
A	370-1	Ta gr2 (100195)	Papillary	no	-
A	810-1	Ta gr2 (031097)	Papillary	no	-
A	1146-1	Ta gr2 (231199)	Papillary	no	-
A	1161-1	Ta gr2 (101299)	Mixed	no	-
A	1006-1	Ta gr2 (231198)	Papillary	no	-
A	942-1	Ta gr2	Papillary	no	24 month.
A	1060-1	Ta gr2	Papillary	no	36 month.
A	1255-1	Ta gr2	Papillary	no	24 month.
B	441-1	Ta gr2	Papillary	no	6 month.
B	780-1	Ta gr2	Papillary	no	2 month.
B	815-2	Ta gr2	Papillary	no	6 month.
B	829-1	Ta gr2	Papillary	no	4 month.
B	861-1	Ta gr2	Papillary	no	4 month.
B	925-1	Ta gr2	Papillary	no	5 month.
B	1008-1	Ta gr2	Papillary	no	5 month.
B	1086-1	Ta gr2	Papillary	no	6 month.
B	1105-1	Ta gr2	Papillary	no	8 month.
B	1145-1	Ta gr2	Papillary	no	4 month.
B	1327-1	Ta gr2	Papillary	no	5 month.
B	1352-1	Ta gr2	Papillary	no	6 month.
B	1379-1	Ta gr2	Papillary	no	5 month.
B	533-1	Ta gr2	Papillary	no	4 month.
B	679-1	Ta gr2	Papillary	no	4 month.
B	692-1	Ta gr2	Papillary	no	5 month.

Group A: Primary tumours from patients with no recurrence of the disease for 2 years.

Group B: Primary tumours from patients with recurrence of the disease within 8 months.

Web Figure C: Number of classification errors vs. number of genes used in cross validation loops.



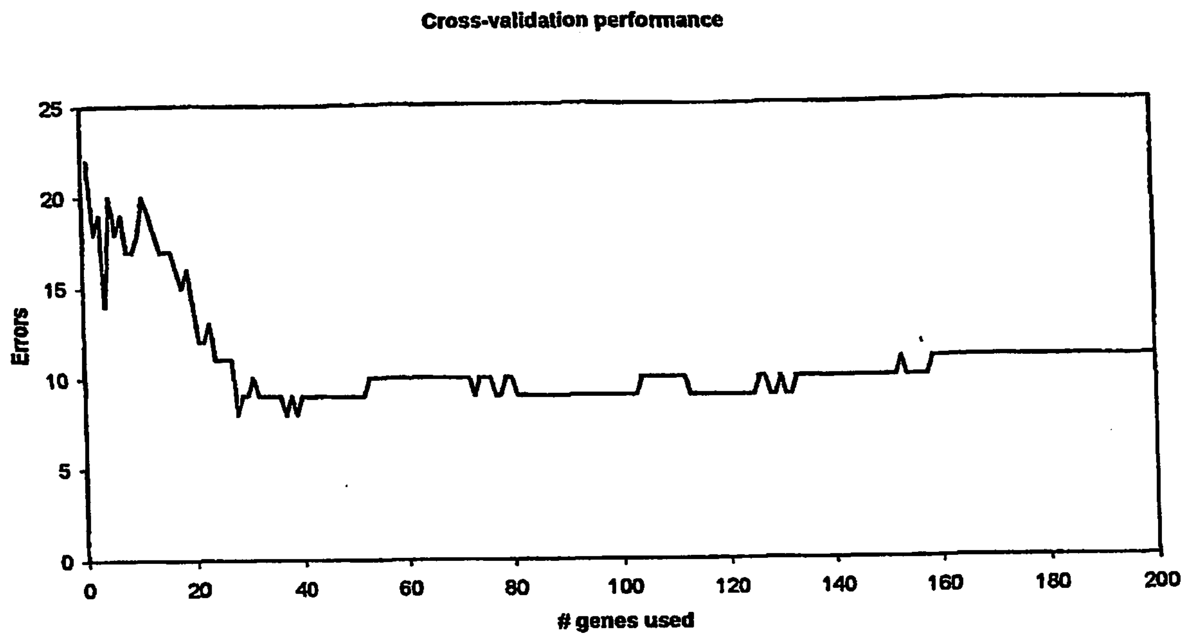
Web Table E: Patient disease course information – recurrence vs. no recurrence

Group	Patient	Tumour (date)	Pattern	Carcinoma *in situ*	Time to recurrence
A	968-1	Ta gr2	Papillary	no	27 month
A	928-1	Ta gr2	Papillary	no	38 month.
A	934-1	Ta gr2 (220798)	Papillary	no	-
A	709-1	Ta gr2 (210798)	Papillary	no	-
A	930-1	Ta gr2 (300698)	Papillary	no	-
A	524-1	Ta gr2 (201095)	Papillary	no	-
A	455-1	Ta gr2 (060695)	Papillary	no	-
A	370-1	Ta gr2 (100195)	Papillary	no	-
A	810-1	Ta gr2 (031097)	Papillary	no	-
A	1146-1	Ta gr2 (231199)	Papillary	no	-
A	1161-1	Ta gr2 (101299)	Mixed	no	-
A	1006-1	Ta gr2 (231198)	Papillary	no	-
A	942-1	Ta gr2	Papillary	no	24 month.
A	1060-1	Ta gr2	Papillary	no	36 month.
A	1255-1	Ta gr2	Papillary	no	24 month.
B	441-1	Ta gr2	Papillary	no	6 month.
B	780-1	Ta gr2	Papillary	no	2 month.
B	815-2	Ta gr2	Papillary	no	6 month.
B	829-1	Ta gr2	Papillary	no	4 month.
B	861-1	Ta gr2	Papillary	no	4 month.
B	925-1	Ta gr2	Papillary	no	5 month.
B	1008-1	Ta gr2	Papillary	no	5 month.
B	1088-1	Ta gr2	Papillary	no	6 month.
B	1105-1	Ta gr2	Papillary	no	8 month.
B	1145-1	Ta gr2	Papillary	no	4 month.
B	1327-1	Ta gr2	Papillary	no	5 month.
B	1352-1	Ta gr2	Papillary	no	6 month.
B	1379-1	Ta gr2	Papillary	no	5 month.
B	533-1	Ta gr2	Papillary	no	4 month.
B	679-1	Ta gr2	Papillary	no	4 month.
B	692-1	Ta gr2	Papillary	no	5 month.

Group A: Primary tumours from patients with no recurrence of the disease for 2 years.

Group B: Primary tumours from patients with recurrence of the disease within 8 months.

Web Figure F: Number of classification errors vs. number of genes used in cross validation loops.



Web Table A: Patient disease course information – class discovery

Group	Patient	Previous tumours	Tumour examined on array	Pattern	Reviewed histology	Subsequent tumours	Carcinoma *in situ**
A	709-1		Ta gr 2 (200297)	Papillary	Ta gr3		no
	968-1		Ta gr 2 (011098)	Papillary	+	Ta gr 2 (150101)	no
	934-1		Ta gr 2 (220798)	Papillary	+		no
	928-1		Ta gr 2 (240698)	Papillary	+		no
	930-1		Ta gr 2 (300698)	Papillary	+		no
B	989-1		Ta gr 3 (281098)	Papillary	+		no
	1264-1		Ta gr 3 (130600)	Papillary	+	Ta gr 2 (231000) Ta gr 2 (220101) Ta gr 2 (300401)	no
	876-5	Ta gr 2 (230398) Ta gr 2 (271098) Ta gr 2 (090699) Ta gr 2 (011199)	Ta gr 3 (170400)	Papillary	+		no
	669-7	Ta gr 2 (101296) Ta gr 2 (150897) Ta gr 1 (161297) Ta gr 3 (270498) Ta gr 2 (220299)	Ta gr 3 (230899)	Papillary	Ta gr2	Ta gr 2 (120100) Ta gr 2 (250500) Ta gr 2 (250900) Ta gr 2 (050201)	no
	716-2	Ta gr 2 (070397)	Ta gr 3 (230497)	Papillary	+	Ta gr 2 (040697) Ta gr 1 (170698)	no
C	1070-1		Ta gr 3 (150399)	Papillary	+	Ta gr 3 (291099)	Subsequent visit
	956-2		Ta gr 3 (061299)	Papillary	+	Ta gr 3 (061200)	Sampling visit
	1062-2		Ta gr 3 (120799)	Papillary	+	T1 gr 3 (161199)	Sampling visit
	1166-1		Ta gr 3 (271099)	Papillary	+		Sampling visit
	1330-1		Ta gr 3 (311000)	Papillary	+		Sampling visit
D	112-10	Ta gr 2 (070794) Ta gr 3 (011294) T1 gr 3(150695) Ta gr 3 (121095) T1 gr 3(040396) Ta gr 2 (200896) Ta gr 2 (111296) Ta gr 2 (230497) Ta gr 2 (030997)	Ta gr 3 (060198)	Papillary	+	Ta gr 3 (110698) T1 gr 3 (191098) Ta gr 3 (240299) T1 gr 3 (050799) T1 gr 3 (081199) T1 gr 3'(180400)	Previous visit
	320-7	T1 gr 3 (011194) T1 gr 3 (150896) Ta gr 3 (100897)	Ta gr 3 (290997)	Papillary	+	Ta gr 3 (290198) Ta gr 3 (290698)	Sampling visit
	747-7	Ta gr 2 (010597) Ta gr 2 (220597) Ta gr 2 (230997) Ta gr 2 (260198) T1 gr 3 (270498) Ta gr 2 (170898)	Ta gr 3 (161298)	Papillary	+	Ta gr 2 (050599) Ta gr 2 (280999) Ta gr 2 (141299)	Sampling visit
	967-3	T1 gr 3 (280998) T1 gr 3 (250199)	Ta gr 3 (140699)	Papillary	+	T1 gr 3 (080999)	Sampling visit
E	625-1		T1 gr 3 (200996)	Papillary	+		No
	847-1		T1 gr 3 (210198)	Papillary	+		No
	1257-1		T1 gr 3 (240500)	Solid	+		Sampling visit
	919-1		T1 gr 3 (220698)	Papillary	+		No
	880-1		T1 gr 3 (300398)	Papillary	+	Ta gr 2 (091198) Ta gr 1 (090399) Ta gr 2 (050900) Ta gr 2 (190301)	No
	812-1		T1 gr 3 (061098)	Papillary	+		No
	1269-1		T1 gr 3 (230600)	Papillary	-		No
	1083-2	Ta gr 2 (280499)	T1 gr 3 (120599)	Papillary	-		No
	1238-1		T1 gr 3 (020500)	Papillary	+	T2 gr 3 (211100) Ta gr 2'(211100)	No
	1065-1		T1 gr 3 (160399)	Papillary	-		Subsequent visit
	1134-1		T1 gr 3 (181099)	Papillary	T2 gr3	T1 gr 3 (280200) T1 gr 3 (020500) T1 gr 3 (131100)	Sampling visit

F	1164-1		T2+ gr 4 (101299)	Solid	gr 3		No
	1032-1		T2+ gr ? (050199)	Mixed	-		Not measured
	1117-1		T2+ gr 3 (010999)	Solid	+		Sampling visit
	1178-1		T2+ gr 3 (200100)	Solid	+		Not measured
	1078-1		T2+ gr 3 (120499)	Solid	+		Not measured
	875-1		T2+ gr 3 (180398)	Solid	+		No
	1044-1		T2+ gr 3 (010299)	Solid	+	T2+ gr 3 (060999)	Not measured
	1133-1		T2+ gr 3 (081099)	Solid	+		Not measured
	1068-1		T2+ gr 3 (220399)	Solid	+		No
	937-1		T2+ gr 3 (280798)	Solid	-		Not measured

Group A: Ta gr2 tumours – no recurrence within 2 years.

Group B: Ta gr3 tumours – no prior T1 tumour and no carcinoma *in situ* in random biopsies.

Group C: Ta gr3 tumours – no prior T1 tumour but carcinoma *in situ* in random biopsies. Group D: Ta gr3 tumours – a prior T1 tumour and carcinoma *in situ* in random biopsies. Group E: T1 gr3 tumours – no prior T2+ tumour. Group F: T2+ tumours gr3/4 – only primary tumours.

* Carcinoma *in situ* detected in selected site biopsies at previous, sampling or subsequent visits.

Web Table B: The 32 genes used in at least 75% (27 times) of the cross validation loops.

Feature	Unigene	Description	Number	Test (B/W)	Testgroup
D83920_at	Hs.252136	ficolin (collagen/fibrinogen domain-containing) 1	31	33.62	3
HG67-HT67_f_at	NA	zinc finger protein SBZF3	35	51.47	1
HG907-HT907_at	Hs.37936	suppressor of variegation 3-9 (Drosophila) homolog 1	35	43.63	1
J05032_at	Hs.80758	aspartyl-tRNA synthetase	35	44.30	1
K01396_at	Hs.297681	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	31	34.24	3
M16591_s_at	Hs.89555	hemopoietic cell kinase	35	38.71	3
M32011_at	Hs.949	neutrophil cytosolic factor 2 (65kD, chronic granulomatous disease, autosomal 2)	35	48.35	3
M33195_at	Hs.743	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide	29	33.12	3
M37033_at	Hs.82212	CD53 antigen	33	34.08	3
M57731_s_at	Hs.75765	GRO2 oncogene	35	37.07	3
M63262_at	NA	Arachidonate 5-lipoxygenase-activating protein	35	37.52	3
S77393_at	Hs.94881	ESTs	35	85.04	2
U01833_at	Hs.81469	nucleotide binding protein 1 (E.coli MinD like)	35	54.81	1
U07231_at	Hs.309763	G-rich RNA sequence binding factor 1	35	80.54	2
U41315_rna1_s_at		ring zinc-finger protein (ZNF127-Xp)	35	89.24	2
U47414_at	Hs.79069	cyclin G2	35	82.49	2
U50708_at	Hs.1265	branched chain keto acid dehydrogenase E1, beta polypeptide (maple syrup urine disease)	35	48.75	1
U52101_at	Hs.9999	epithelial membrane protein 3	34	34.39	3
U74324_at	Hs.90875	RAB interacting factor	35	47.87	1
U77970_at	NA	neuronal PAS domain protein 2 (NPAS2)	30	72.77	2
U90549_at	Hs.236774	high-mobility group (nonhistone chromosomal) protein 17-like 3	35	48.41	1
X13334_at	Hs.75627	CD14 antigen	34	35.00	3
X54489_rna1_at	NA	melanoma growth stimulatory activity	34	75.37	2
X57579_s_at	Hs.727	inhibin, beta A (activin A, activin AB alpha polypeptide)	35	89.41	2
X64072_s_at	Hs.83968	integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit)	35	40.08	3
X68194_at	Hs.80919	synaptophysin-like protein	29	72.29	2
X73882_at	Hs.146388	microtubule-associated protein 7	35	89.29	2
X78520_at	Hs.174139	chloride channel 3	35	83.36	2
X95632_s_at	Hs.343575	abl-interactor 12 (SH3-containing protein)	33	41.11	1
Z29331_at	Hs.28505	ubiquitin-conjugating enzyme E2H (homologous to yeast UBC8)	35	63.45	1
Z48605_at	Hs.5123	inorganic pyrophosphatase	29	72.12	2
Z74615_at	Hs.172928	collagen, type I, alpha 1	35	108.84	2

Feature: Accession number on HuGeneFL array.

Number: Number of times used in cross validation.

Testgroup: genes selected from having a high value of B/W when comparing Ta with T1 (1), Ta with T2 (2), and T1 with T2 (3).

Test (B/W): To test the class separation performance of the 32 selected genes we compared their B/W ratios with the similar ratios of all the genes calculated from permutations of the arrays. For each permutation we construct three pseudogroups, pseudo-Ta, pseudo-T1, and pseudo-T2, so that the proportion of samples from the three original groups is approximately the same in the three pseudogroups. We then calculated the three B/W ratios, B(Ta/T1)/W, B(Ta/T2)/W, and B(T1/T2)/W, based on the pseudogroups and selected the 32 highest values in the same way as for the actual data. For the highest scoring gene among the 32 selected we found that 500 values obtained from the permutations have a mean value of 19.04 with the highest observed being 43.91. This should be compared to the value 108.84 from the actual data in Table 4. For the lowest scoring gene we found that the 500 values had a mean value of 9.69 with the highest being 20.55 (to be compared with 33.12 from the table).

Web Table E: Patient disease course information – recurrence vs. no recurrence

Group	Patient	Tumour (date)	Pattern	Carcinoma *in situ*	Time to recurrence
A	968-1	Ta gr2	Papillary	no	27 month
A	928-1	Ta gr2	Papillary	no	38 month.
A	934-1	Ta gr2 (220798)	Papillary	no	-
A	709-1	Ta gr2 (210798)	Papillary	no	-
A	930-1	Ta gr2 (300698)	Papillary	no	-
A	524-1	Ta gr2 (201095)	Papillary	no	-
A	455-1	Ta gr2 (060695)	Papillary	no	-
A	370-1	Ta gr2 (100195)	Papillary	no	-
A	810-1	Ta gr2 (031097)	Papillary	no	-
A	1146-1	Ta gr2 (231199)	Papillary	no	-
A	1161-1	Ta gr2 (101299)	Mixed	no	-
A	1006-1	Ta gr2 (231198)	Papillary	no	-
A	942-1	Ta gr2	Papillary	no	24 month.
A	1060-1	Ta gr2	Papillary	no	36 month.
A	1255-1	Ta gr2	Papillary	no	24 month.
B	441-1	Ta gr2	Papillary	no	6 month.
B	780-1	Ta gr2	Papillary	no	2 month.
B	815-2	Ta gr2	Papillary	no	6 month.
B	829-1	Ta gr2	Papillary	no	4 month.
B	861-1	Ta gr2	Papillary	no	4 month.
B	925-1	Ta gr2	Papillary	no	5 month.
B	1008-1	Ta gr2	Papillary	no	5 month.
B	1086-1	Ta gr2	Papillary	no	6 month.
B	1105-1	Ta gr2	Papillary	no	8 month.
B	1145-1	Ta gr2	Papillary	no	4 month.
B	1327-1	Ta gr2	Papillary	no	5 month.
B	1352-1	Ta gr2	Papillary	no	6 month.
B	1379-1	Ta gr2	Papillary	no	5 month.
B	533-1	Ta gr2	Papillary	no	4 month.
B	679-1	Ta gr2	Papillary	no	4 month.
B	692-1	Ta gr2	Papillary	no	5 month.

Group A: Primary tumours from patients with no recurrence of the disease for 2 years.

Group B: Primary tumours from patients with recurrence of the disease within 8 months.

Web Table F: Recurrence prediction results of 39 gene cross-validation loops.

Group	Patient	Tumour (date)	Prediction	Error	Prediction strength
A	968-1	Ta gr2	0		0.19
A	928-1	Ta gr2	0		0.49
A	934-1	Ta gr2 (220798)	0		1.73
A	709-1	Ta gr2 (210798)	0		0.45
A	930-1	Ta gr2 (300698)	0		0.82
A	524-1	Ta gr2 (201095)	0		0.14
A	455-1	Ta gr2 (060695)	1	*	0.68
A	370-1	Ta gr2 (100195)	0		0.32
A	810-1	Ta gr2 (031097)	0		0.45
A	1146-1	Ta gr2 (231199)	0		0.98
A	1161-1	Ta gr2 (101299)	0		0.03
A	1006-1	Ta gr2 (231198)	1	*	1.57
A	942-1	Ta gr2	0		0.31
A	1060-1	Ta gr2	1	*	0.81
A	1255-1	Ta gr2	1	*	0.71
B	441-1	Ta gr2	1		1.03
B	780-1	Ta gr2	1		0.37
B	815-2	Ta gr2	1		0.35
B	829-1	Ta gr2	1		0.75
B	861-1	Ta gr2	0	*	2.55
B	925-1	Ta gr2	1		0.78
B	1008-1	Ta gr2	0	*	0.12
B	1086-1	Ta gr2	0	*	0.51
B	1105-1	Ta gr2	1		0.37
B	1145-1	Ta gr2	1		0.44
B	1327-1	Ta gr2	1		1.96
B	1352-1	Ta gr2	0	*	0.97
B	1379-1	Ta gr2	1		0.67
B	533-1	Ta gr2	1		0.31
B	679-1	Ta gr2	1		0.82
B	692-1	Ta gr2	1		0.45

Group A: Primary tumours from patients with no recurrence of the disease for 2 years.
Group B: Primary tumours from patients with recurrence of the disease within 8 months.
Prediction: 0=no recurrence, 1=recurrence.

Prediction strength: The relative difference between the distance to the closest and the second closest group compared to the distance to the closest group.

Web Table G: The 26 genes used in at least 75% (29 times) of the cross validation loops.

Feature	Unigene	Description	Number	Test (W-N)
AF006041_at	Hs.336916	death-associated protein 6	31	0.054 (161-7)
D21337_at	Hs.408	collagen, type IV, alpha 6	31	0.058 (160-6)
D49387_at	-	NADP dependent leukotriene b4 12-hydroxydehydrogenase	31	0.118 (313-8)
D64154_at	Hs.90107	cell membrane glycoprotein, 110000M(r) (surface antigen)	31	0.078 (165-9)
D83780_at	Hs.8294	KIAA0196 gene product	31	0.094 (159-4)
D87258_at	Hs.75111	protease, serine, 11 (IGF binding)	30	0.112 (168-11)
D87437_at	Hs.15087	chromosome 1 open reading frame 16	31	0.058 (160-6)
HG1879-HT1919_at	-	Ras-Like Protein Tc10	31	0.122 (314-7)
HG3076-HT3238_s_at	-	Heterogeneous Nuclear Ribonucleoprotein K, Alt. Splice 1	31	0.080 (309-17)
HG511-HT511_at	-	Ras Inhibitor Inf	31	0.348 (319-2)
L34155_at	Hs.83450	laminin, alpha 3	31	0.122 (314-7)
L38928_at	Hs.118131	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)	29	0.348 (319-2)
L49169_at	Hs.75678	FBJ murine osteosarcoma viral oncogene homolog B	31	0.108 (155-2)
M16938_s_at	Hs.820	homeo box C6	29	0.09 (170-16)
M63175_at	Hs.80731	autocrine motility factor receptor	29	0.098 (308-18)
M64572_at	Hs.153932	protein tyrosine phosphatase, non-receptor type 3	31	0.064 (305-31)
M98528_at	Hs.79404	neuron-specific protein	31	0.122 (314-7)
U21858_at	Hs.60679	TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 32 kD	31	0.122 (314-7)
U45973_at	Hs.178347	SKIP for skeletal muscle and kidney enriched inositol phosphatase	31	0.094 (310-14)
U58516_at	Hs.3745	milk fat globule-EGF factor 8 protein	29	0.100 (175-28)
U62015_at	Hs.8867	cysteine-rich, angiogenic inducer, 61	31	0.108 (169-13)
U66702_at	Hs.74624	protein tyrosine phosphatase, receptor type, N polypeptide 2	31	0.146 (149-1)
U70439_s_at	Hs.84264	acidic protein rich in leucines	30	0.08 (309-17)
U94855_at	Hs.7811	eukaryotic translation initiation factor 3, subunit 5	30	0.092 (311-12)
X63469_at	Hs.77100	general transcription factor IIE, polypeptide 2	31	0.092 (311-12)
Z23064_at	Hs.146381	RNA binding motif protein, X chromosome	30	0.066 (307-24)

Feature: Accession number on HuGeneFL array.
Number: Number of times the gene has been used in a cross-validation loop.
Test: The numbers in parenthesis are the value W of the Wilcoxon test statistic for no difference between the two groups together with the number N of genes for which the Wilcoxon test statistic is bigger than or equal to the value W. The test value is obtained from 500 permutations of the arrays. In each permutation we form new pseudogroups where both of the pseudogroups have the same proportion of arrays from the two original groups. For each permutation we count the number of genes for which the Wilcoxon test statistic based on the pseudogroups is bigger than or equal to W, and the test value is the proportion of the permutations for which this number is bigger than or equal to N. Thus the test value measures the significance of the observed value W. Consequently, for most of our selected genes we only find as least as good predictive genes in about 10% of the formed pseudogroups.